THIS WEEK

EDITORIALS

MOONSHOT Lunar scientists ask public to stump up US$1 billion for mission **p.466**

WORLD VIEW We must ease the rising burden on peer reviewers **p.467**

FISH SUPPER Seals home in on noisy research tags to find prey **p.469**

Agree to agree

The US–China emissions agreement raises hopes for international cooperation on a climate accord. But it does not go far enough.

Roughly 44% of the carbon dioxide that humans release into the atmosphere each year comes from China and the United States. These countries are the big fish at the United Nations climate negotiations, and for years they have been at loggerheads, each deflecting calls to curb emissions by pointing to the other. As talks languished and emissions increased, the rest of the world's major emitters often seemed content to sit back and point the finger at them. Now that these two powerhouse polluters have brokered an unexpected deal on emissions, can the world hope that those days are in the past? All eyes are on the UN climate meeting in Lima next week (see page 473).

The chances of forming a meaningful climate agreement at the follow-up UN summit in Paris next year have clearly improved, but a dose of scepticism is warranted. Under the US–China agreement, inked in Beijing on 12 November by presidents Barack Obama and Xi Jinping, the United States would curb its emissions by at least 26% by 2025, and China would hasten the development of low-carbon energy to ensure that its own soaring emissions peak around 2030. These are not insignificant targets, but both nations could do more if they were really serious about addressing climate change. Moreover, both pledges come shrouded in their own particular doubt.

From the US perspective, Obama has two years left in his presidential term to get the ball rolling and even if he succeeds, that ball will roll into an uncertain future. His primary weapon is a proposed set of regulations for existing power plants, which the US Environmental Protection Agency says would reduce emissions by 30% from 2005 levels by 2030. Assuming that they clear the inevitable court challenges, these regulations and the action already taken on vehicle fuel efficiency will go a long way, but more will be needed.

As it stands, many Republicans are lined up against the president's climate policies. They have already been critical of Obama's agreement with China, and as of January they will control both houses of Congress. Some observers predict that policies on climate could be among the major issues in the presidential election two years from now. That would be a welcome first, because much will ride on the outcome.

As for China, the headline promise is maddeningly vague: 'around 2030' does not tell us precisely when emissions will peak, and Xi did not specify how high emissions will climb before then. China is already on track to meet its existing goal of producing 15% of its power from low-carbon sources by 2020. So its promise to extend that to 20% by 2030 makes the latest commitment less than revolutionary. And although some energy researchers have suggested that China could level off its emissions by 2025, most baseline scenarios suggest that without active engagement, the country's emissions would continue to rise until 2050, albeit slowing down once more Chinese citizens have finished filling their homes with energy-hungry appliances.

This agreement has as much to do with political momentum as commitments. The stand-off between the United States and China is emblematic of a larger rift in the negotiations and has its roots in both morality and science. Developing countries rightly expect those who have profited from polluting the atmosphere to lead the way in curbing emissions; industrialized countries rightly counter that they cannot do it alone, given that most of the growth in greenhouse-gas emissions is in the developing world, where more than 1 billion people still live without electricity. Unfortunately, the climate does not care about such questions.

"The stand-off between the United States and China is emblematic of a larger rift."

Five years ago, at the most recent headline climate summit in Copenhagen, world leaders took their first step in breaking down the legal wall between developed and developing countries. Until then, under the 1997 Kyoto Protocol, only developed countries — notably minus the United States, which ducked out — had obligations to reduce emissions. In Copenhagen, many developing countries stepped forward with climate pledges, but the negotiations nearly collapsed. Significant battles remain over commitments, financial aid and how to structure an agreement, but most countries now accept that this must be a collective effort.

In theory, the US–China agreement is the last major piece of this puzzle. If it translates into cooperation on a new climate accord, other countries may be encouraged to engage seriously. At a minimum, those who have been pointing the finger at the United States and China would need to come up with another excuse.

All involved will get the first glimpse of how this changes the international dynamic when negotiators gather in Lima. Fingers crossed. ■

Ebola opportunity

A slowdown in new cases offers a chance for control efforts to get ahead of the epidemic.

An apparent slowdown in new cases of Ebola disease in Liberia and Guinea should be taken advantage of. Almost one year after an Ebola epidemic began in West Africa there are at last encouraging signs that it may be receding in some regions. But those responding to the epidemic must not drop their guard — rather, they should seize upon the chance to finish the job.

"Today, we — two dumbfounded doctors — stare at our empty blackboard. We have no more patients." Last week, that declaration was blogged by a doctor with the humanitarian agency Médecins Sans Frontières (MSF), also known as Doctors Without Borders, at an Ebola treatment centre in the Foya region of Liberia. It is the same

story in many parts of the country: empty beds that would have been unthinkable just a few weeks ago when Ebola treatment centres were overflowing. Nationally, the growth in the numbers of those infected in Liberia, the worst-affected country, is no longer exponential but has flattened off.

The epidemic has also stabilized in Guinea. But a resurgence of cases in Sierra Leone is a timely reminder that until Ebola is eliminated throughout West Africa, it remains a major threat. As of 18 November, Ebola has infected at least 15,000 people and killed 5,440 of them in these three main affected countries. But the worst-case scenarios predicted by mathematical modellers, which projected a steady apocalyptic rise in Ebola case numbers, have proved far off the mark (see *Nature* **515,** 18; 2014).

Although complacency is as unwise as it is hopefully unlikely — a lull in Ebola cases in the spring prompted authorities to drop their guard, only to see the virus return with a vengeance — there are reasons to believe that the current lull in Liberia and Guinea may continue. And that offers an opportunity to roll back the epidemic at last.

The exact causes of the lull are unclear. Belated international Ebola control efforts are only now beginning to kick in, and have no doubt contributed. But much of the slowdown is perhaps due to Africans themselves coming to terms with the epidemic and blocking its main routes of transmission. In particular, there has been a reduction in traditional burial practices, which are a key source of spread.

The slowing of new cases in Liberia and Guinea is a welcome reprieve for the health-care workers and scientists who have toiled to control a virus that for months has held the advantage. It is an opportunity to regroup, to consolidate gains, and to go all the more on the offensive.

Until recently, MSF, based in Geneva, Switzerland, was the only serious international presence fighting Ebola on the ground, but logistics meant that it could operate only a few large centralized treatment centres. These large centres, often with hundreds of beds, are still needed to absorb any resurgence, particularly in urban areas. But having only large centres is not ideal. Patients often have to travel for many hours or even days to reach them, and by the time they make it are often beyond recovery. They are also likely to have contaminated others en route, so fuelling the spread of the virus.

With its caseloads falling in recent weeks, MSF is coming out of the trenches and taking the fight to the virus, sending mobile teams and smaller treatment centres to the sites of new outbreaks to try to nip them in the bud. MSF sensibly wants other aid groups to adapt in a similar way. It will be a challenge for the more bureaucratic UN Mission for Ebola Emergency Response, and the US and other national Ebola-treatment efforts, to quickly change their plans, because they are mainly based around large centres. But it is crucial that the response to Ebola is flexible in the face of the shifting epidemiology.

"Those responding to the epidemic must not drop their guard."

The slowdown is also buying precious time for the testing of drugs and vaccines: clinical trials of vaccines in particular are being fast-tracked, with the first results due at the end of 2014. Unfortunately, however, drugs and vaccines have captured the spotlight and resources, while more mundane interventions that could have an immediate impact have been neglected. Better rehydration and electrolyte control can dramatically reduce mortality: the case fatality rate for patients treated in rich countries has been a fraction of the 70% seen in West Africa. Testing convalescent blood and serum from survivors — a potentially game-changing treatment — should also be a priority.

At the start of October, the United Nations and the World Health Organization set quantitative targets for safe burials, contact tracing and other key public-health control measures, which the international community was to meet by 1 December. It is already obvious that most of these targets will not be met. The breathing space offered by the current lull in Liberia and Guinea offers an opportunity to fill gaps and ramp up coverage of countermeasures. It must not be wasted. ■

Moon on a stick

A crowdfunded lunar mission might seem like a long shot — but there is no harm in trying.

The crowdfunding platform Kickstarter is popular with inventors of fashionable bike helmets, hover boards and even a smart frying pan that tells you when to flip a steak. But last week the site that has funded thousands of films, games and gadgets launched a funding effort for something much bigger: a mission to the Moon.

On 19 November, under the banner of 'Lunar Mission One', a UK-led consortium announced a goal to put a lander on the Moon by 2024 and to retrieve and analyse samples from 100 metres below the lunar south pole. The mission itself would cost around US$1 billion. For starters, it needs $1 million by 17 December. As *Nature* went to press, it has more than half of that.

Attempting to invert the fund-raising model for science missions, the project would get its cash by encouraging many thousands of people to give a few dollars. In return, investors get the chance to preserve a little bit of themselves in a time capsule that will fill the borehole: either in a digital form, in a 'memory box', or with a strand of hair. The latter would cost as little as $80.

This is not the first science project to seek crowdfunding. For example, last year, synthetic biologist Omri Amirav-Drory, founder of the company Genome Compiler, raised $480,000 on Kickstarter to create glowing plants. Nor is it the first private venture to shoot for the Moon — the companies that compete for the Google Lunar X Prize are the most notable.

So how seriously should we take the new Moon shot? Certainly the institutions involved are solid enough. University College London and RAL Space, part of the UK Science and Technology Facilities Council and a partner in more than 200 space missions, have assessed the feasibility of the mission. The Open University in Milton Keynes, UK, is working on the educational side. High-profile celebrity scientists and former UK science ministers have clamoured to back the venture, and seasoned academics from universities across the United Kingdom have built the mission's science case.

The promised science is interesting. Europe has never sent a lander to the lunar surface, and no nation has ever visited its south pole. Permanently shadowed craters there are thought to contain water, and digging deep into the surface could answer countless unsolved questions about the Moon's history. And the timing is good. Space science is basking in the glow of ESA's Rosetta mission, which landed a probe on a comet earlier this month.

But $1 billion? Organizers aim to fill a gap in public funding in a way that neither detracts from existing space missions nor puts governments in a predicament. The mission gets funding only if people care enough to contribute, says Richard Holdaway, director of RAL Space. "It's not about deciding whether to spend money on a space programme or a new hospital," he says. "It's democracy at its greatest level."

Perhaps wary of pouring cold water on an aspirational and ambitious plan, sceptics have been surprisingly hard to find. The effort is plainly ambitious. But, the message seems to be, where is the harm in trying? If Lunar Mission One misses its funding target, the programme simply stops, having broken the first rule of any sales effort — offer a product that enough consumers want. ■

NATURE.COM
To comment online, click on Editorials at:
go.nature.com/xhunqv

Open access is tiring out peer reviewers

As numbers of published articles rise, the scholarly review system must adapt to avoid unmanageable burdens and slipping standards, says **Martijn Arns**.

Scientists like to complain about peer review. No researcher wants to be told that their work is flawed, unworthy or just plain wrong. But in recent months, I received reviews of my own submitted papers that suggest reviewers simply did not read the manuscript properly.

This is not nitpicking over matters of opinion or interpretation. In one instance, a reviewer complimented the double-blind placebo-controlled nature of our study, and made methodological comments related to that. Yet the study was not placebo controlled. In fact, participants were randomly assigned to three different active treatments. That is a serious mistake and undermines the supposed internal quality control of the peer-review system.

Conversations with colleagues reveal similar concerns about peer-review quality, and suggest that the scale of the problem has increased over the past few years. These are anecdotal reports, but they do raise a serious question: as the number of academic papers and scientific journals published continues to grow, can the peer-review system cope?

The migration of scholarly journals from print to digital increases the burden on reviewers. Online publications have no page budgets or print costs, and so can publish as much as they like. Once, this process was managed by editors who would decide whether to send a paper out for review, or to simply reject it. This system had its own disadvantages but it seemed to keep the total number of papers that required review at a manageable level. The default option for many online journals seems to be to send all submissions out for review. The rise of the open-access (OA) movement compounds this effect. The business case for online OA journals, to which authors pay submission fees, works best at high volume. And for many of these journals, submitted work is published as long as it is methodologically sound. It does not have to demonstrate, for example, the novelty or societal relevance that some traditional journals demand.

The OA publisher Frontiers, for example, focuses on: "certifying the accuracy and validity of articles, not on evaluating their significance".

I think that some reviewers take the removal of the need for significance as a signal that they need to read and evaluate only the methods and statistics sections of a paper under review, and pay less attention to its rationale and wider context. One positive development of this is that papers that are important, but of limited interest, can get published, such as 'null' findings and failed replications. But given the 'publish or perish' nature of modern research, if scientists can publish more papers, they will. In this way, OA and other online journals both meet and create the demand for a massive rise in academic output. The OA journal *PLoS ONE*, for example, has published more than 105,000 papers since 2006, and Frontiers more than 20,000 since 2007. If at least two reviewers saw each manuscript, that amounts to more than 250,000 reviews for those two publishers alone.

If the number of journals and manuscripts grows faster than the number of scientists, the pressure on peer reviewers has to increase. Is that happening? It is hard to find reliable data. The annual number of articles indexed in the publisher Elsevier's Scopus database increased from around 1.2 million in 2000 to roughly 2.7 million in 2013. That is an increase of 113%, but some of this rise is simply due to articles from more journals being included in the later count. Available figures suggest that the increase in scientists is slower: 2.8% per year in the European Union (between 2006 and 2011) and just 1.5% in the United States, but it is harder to track the faster rates of change in countries such as China. A 2014 survey of 3,000 scientists by Elsevier found that only 29% complained that pressure is increasing on reviewers — but that figure is 10% higher than in 2009.

INCREASED PRESSURE MEANS **PAPERS** ARE ASSIGNED TO **REVIEWERS** WHO ARE **NOT EXPERTS** IN THE AREA.

One result of increased pressure is that papers are assigned to reviewers who are not experts in the area. They might have the technical ability to evaluate methods and results sections — as these OA journals require — but lack the expertise to evaluate a full paper, including introduction and discussion. This matters. Reviewers should verify that authors are quoting the right literature to support their rationale. Citing obsolete studies will set back science, because invalid conclusions might be kept alive.

To protect quality reviewing, a hybrid model should be considered. I suggest a two-tier system, in which some papers are not reviewed before publication at all and are instead subject to a post-publication peer review. Some manuscripts are of interest mainly to scientists, such as null findings, methodological studies or straight repeats of previous experiments. There is great value in publishing these papers, but perhaps not in sending them all out for review. This would free up peer reviewers to focus on papers with more direct societal impact, where the question of whether to publish at all is more relevant. Pre-publication review is more important there, because it protects the lay audience from being exposed to 'miracle cures' and wild claims.

In my view, we must look at the massive expansion of online publications (most of which are OA journals) as a disruptive technology, resulting in overworked and fatigued reviewers. Quality will suffer — across the board — unless something is done. ■

➲ **NATURE.COM**
Discuss this article online at:
go.nature.com/rst2fv

Martijn Arns *is director and researcher at Research Institute Brainclinics in Nijmegen, the Netherlands.*
e-mail: martijn@brainclinics.com

RESEARCH HIGHLIGHTS

Selections from the scientific literature

MATERIALS

Diodes printed in three dimensions

Researchers have created a light-emitting diode (LED) by three-dimensional (3D) printing of five different materials — expanding the number and type of material that can be printed in this way.

This technique involves depositing materials layer by layer until a 3D object is formed. Michael McAlpine and his colleagues at Princeton University in New Jersey used the technology to print a millimetre-sized LED based on quantum dots — nanoscale crystals that emit light.

They seamlessly printed an organic polymer and an indium and gallium metal for the electrodes; a silver metallic interconnect; a layer of quantum dots; and a conductive plastic layer. The entire device was printed onto a curved contact lens.

Other devices, including solar cells and transistors, could be made in this way, the researchers say.

Nano Lett. **http://dx.doi.org/10.1021/nl5033292 (2014)**

IMAGINECHINA/CORBIS

GEOLOGY

Earthquake risk for North China city

Tianjin (pictured), a Chinese city of 11 million people not far from Beijing, lies atop a seismic fault that could be overdue for a large earthquake.

An Yin of the University of California in Los Angeles and his colleagues analysed modern and historical records of earthquakes in northern China. Mapping their locations revealed a 160-kilometre-long fault segment running through Tianjin, roughly 100 kilometres southeast of Beijing. This area has not experienced a major tremor for about 8,400 years, and the authors estimate that a quake of roughly magnitude 7.5 is either overdue or will strike in the next 2,000 to 3,000 years.

Given the region's complex fault structure, however, other factors could explain the lack of major earthquakes, such as multiple smaller quakes releasing energy from the fault.

Geology **http://doi.org/w8j (2014)**

ZOOLOGY

Termite eggs ward off sperm

Ageing termite queens produce new queens asexually by laying eggs without any openings that normally allow sperm to pass through.

KENJI MATSUURA

In termite colonies, queens can reproduce both asexually to generate new queens and sexually to produce other colony members. Toshihisa Yashiro and Kenji Matsuura at Kyoto University in Japan analysed eggs collected from field colonies of the termite *Reticulitermes speratus* (**pictured**). They found that in the eggs that had no openings for sperm, the embryos developed without any genetic contribution from the male. Eggs from older queens tended to have few or no openings compared with eggs from younger queens.

This is a rare example of a female animal controlling the fertilization of her eggs even when males are present, the authors say.

Proc. Natl Acad. Sci. USA **http://doi.org/w87 (2014)**

CHEMISTRY

Rapid synthesis a thousand times

A robotic system that can carry out and analyse 1,536 chemical reactions in less than a day could help to accelerate drug discovery.

Tim Cernak and Spencer Dreher at pharmaceutical company Merck in Massachusetts and New Jersey and their colleagues used the system to couple a model substrate with 16 other molecules, in combination with 16 different palladium catalysts and 6 basic reagents. Each of these unique mixtures was dissolved in 1 microlitre of solvent, automatically dispensed into a separate chamber on a 1,536-well plate, and its products were analysed to determine optimum reaction conditions.

The researchers also coupled pairs of more-complex, drug-like substrates, and easily scaled up successful reactions by 1,000 times or more. This high-throughput approach could rapidly assess synthetic routes to a wide range of drug candidates without wasting precious starting materials.

Science **http://doi.org/w9d (2014)**

MICROBIOLOGY

RNA pockets help parasites to infect

Parasitic worms release tiny sacs filled with small RNAs that disable immune responses in infected mice.

The membrane-bound sacs, or exosomes, sprout from cells and contain proteins and nucleic acids. Amy Buck at the University of Edinburgh, UK, and her team found that the nematode *Heligmosomoides polygyrus*, which infects the mouse gut, produces exosomes containing microRNAs (miRNAs) and 'Y RNAs' that can affect gene expression. The exosomes also carried a protein required to process those RNAs.

Mice exposed to exosomes showed a reduced immune response to an allergen compared to unexposed mice. The sacs also lowered expression of some immune-related genes in mouse cells in a lab dish. Moreover, the animals had miRNA from another worm called *Litomosoides sigmodontis* in their blood, suggesting that miRNA is secreted by other nematodes.

Nature Commun. 5, **5488 (2014)**

GEOPHYSICS

How Greenland got its ice

Changes in Earth's mantle and crust allowed Greenland to accumulate its massive ice sheet over the past few million years.

Bernhard Steinberger of the German Research Centre for Geosciences in Potsdam and his colleagues used various models to reconstruct past plate-tectonic activity. They found that pulses of molten rock from deep within Earth rose up and thinned the overlying crust, which led to an uplift of eastern Greenland by more than 3 kilometres above sea level. Then, two rotations of the crust carrying Greenland shifted it 18° farther north in latitude. Once high and northerly enough, Greenland could begin accumulating ice year-round.

The discovery shows how changes deep inside Earth can drive environmental changes on the surface.

Terra Nova **http://doi.org/w8q (2014)**

MATERIALS

Blu-ray patterns pump up solar cells

The surface pattern on a Blu-ray disc can be used to boost solar-cell performance.

Light is absorbed and scattered in unusual ways by nanometre-scale patterns found on iridescent surfaces, such as insect wings, because the patterns are neither completely periodic nor random. They also allow solar cells to absorb more light, but making such patterns in photonic devices is difficult and expensive. Cheng Sun, Jiaxing Huang and their colleagues at Northwestern University in Evanston, Illinois, discovered that the pits and islands on the surface of Blu-ray movie discs have the same pattern. They used these discs to imprint the patterns on to an organic thin film of a solar cell.

The device absorbed 22% more of the energy from incoming sunlight than an unpatterned solar cell.

Nature Commun. 5, **5517 (2014)**

NEUROSCIENCE

Epilepsy controlled from a distance

Disrupting electrical activity in a brain region not directly affected by epilepsy could be a way to control treatment-resistant forms of the disorder.

Esther Krook-Magnuson and her colleagues at the University of California, Irvine, mimicked epilepsy in mice by injecting a chemical into the hippocampus, where seizures arise in a common form of the human disease that is hard to treat. The mice had been genetically modified so that electrical activity in their brains could be controlled with light.

When the team excited or inhibited neurons in the mouse cerebellum, spontaneous seizures recorded in the hippocampus became shorter. When they excited neurons in the midline cerebellum, the seizures also became less frequent. Altering the activity of neurons in the hippocampus, however, had no effect on seizure frequency.

eNeuro **http://dx.doi.org/10.1523/eneuro.0005-14.2014 (2014)**

SOCIAL SELECTION

Popular articles on social media

Old papers find new life online

Search engines have revolutionized how scientists find papers — especially articles that have been around for a while. A team of researchers at Google has documented a surge in the citation rate for older papers. The study found that 36% of citations in 2013 were to papers that were at least 10 years old — a 28% increase since 1990. Scientists had a range of responses online. Carlos Baquero, a computer scientist at the University of Minho in Braga, Portugal, tweeted: "Older articles are now more accessible and thus their impact has grown. Knowledge escapes tyranny of time."

The authors say that the digitization of journal archives and online search engines have made it easier than ever to find older papers.

Preprint at http://arxiv.org/abs/1411.0275 (2014)

Based on data from altmetric.com. Altmetric is supported by Macmillan Science and Education, which owns Nature Publishing Group.

NATURE.COM
For more on popular papers:
go.nature.com/xjushf

AMANDA STANSBURY/UNIV. ST ANDREWS

ANIMAL BEHAVIOUR

Fish tags guide seal predators

Seals can home in on acoustic tags routinely attached to fish by marine scientists.

These small, sound-emitting devices are often used to track fish populations. Vincent Janik at the University of St Andrews, UK, and his colleagues allowed 10 captive grey seals (*Halichoerus grypus*; **pictured**) to explore 20 boxes in a pool. One box was baited with an untagged fish, another contained a tagged fish and the rest were empty.

In a series of tests, the seals found the tagged fish with increasing speed, and homed in on it faster than on the untagged fish. In later experiments in which no fish bait was used, the animals still generally visited tagged boxes faster than untagged boxes.

This adds to evidence that marine mammals can use human-generated sounds to find prey in the wild.

Proc. R. Soc. B 282, **20141595 (2014)**

NATURE.COM
For the latest research published by *Nature* visit:
www.nature.com/latestresearch

SEVEN DAYS *The news in brief*

POLICY

Oil-pipeline vote

On 18 November, the US Senate rejected legislation to approve the controversial Keystone XL pipeline, which would connect oil sands in Alberta, Canada, to US refineries near the Gulf of Mexico. Supporters fell one vote shy of the 60 needed to advance the pipeline proposal in the Senate; the House of Representatives had approved the bill on 14 November. The issue is expected to resurface in January, when Republicans who favour the pipeline take control of the US Congress.

Climate pledges

At a conference for the Green Climate Fund in Berlin last week, 21 donor countries promised US$9.3 billion to help developing nations to deal with climate change. The total fell just short of the $10 billion that the fund, established in 2011, was meant to have held by the end of 2014.

Clinical-trials data

The US National Institutes of Health and the Food and Drug Administration proposed regulations on 19 November that would make it harder for researchers and companies to hide negative results and harmful side effects that occur in clinical trials. See page 477 for more.

Runaway emissions

Without a significant new international climate agreement, greenhouse-gas emissions are set to overshoot targets intended to limit global warming to 2 °C above pre-industrial levels, the United Nations Environment Programme reported on 19 November. On the basis of existing climate commitments, the agency projects that annual global emissions will exceed targets by as much as 23% in 2020 and 40% in 2030. The report suggests that for the goal to be met, net carbon dioxide emissions must fall to zero between 2055 and 2070, followed by the same reduction in other greenhouse gases.

Polar code

Ships operating in polar waters will be regulated internationally for the first time under a 'polar code' adopted by the International Maritime Organization on 21 November. The United Nations' shipping agency agreed the code, which will regulate the types of ship that are allowed to operate in Arctic and Antarctic waters and how they are run, in the wake of high-profile accidents in Antarctica and an increasingly ice-free Arctic. It aims to safeguard both human life and the delicate environments around the poles. See go.nature.com/jd5z4t for more.

BUSINESS

Charity windfall

The Cystic Fibrosis Foundation has made US$3.3 billion by selling its rights to royalties on cystic-fibrosis drugs, it announced on 19 November. The charity, based in Bethesda, Maryland, received the rights as a condition of its early $75-million investment in cystic-fibrosis therapies developed by Vertex Pharmaceuticals of Boston, Massachusetts. The high price of the resulting drugs — about $300,000 for a year of treatment — has caused controversy (see go.nature.com/noodlm). The foundation sold the rights to Royalty Pharma of New York City.

Drug-cost shock

It costs US$2.56 billion to develop a drug that makes it to market, the Tufts Center for the Study of Drug Development reported on 18 November. The figure is more than double its 2003 estimate. The enormous cost attracted immediate criticism

ASHLEY COOPER/CORBIS

Wave power hits choppy waters

Pelamis Wave Power, based in Edinburgh, UK, said on 21 November that it lacked funds to continue development of its system to generate electricity from oceanic wave power. The company will enter administration, a rescue mechanism that allows it to keep operating and assess options for its future. Pelamis's technology (pictured) uses connected buoys atop the ocean that wriggle in a snake-like motion. Hydraulic pumps along the chain use the motion to drive fluids to an on-board generator. The firm, which has been testing a pair of 750-kilowatt machines in the Orkney Islands, lost a partnership with German utilities firm E.ON last year. Pelamis and other wave-power firms have had trouble attracting commercial interest from the energy sector, in part because current devices cannot withstand battering by the seas over the long term (see *Nature* **508,** 302–304; 2014).

from charities and non-governmental organizations involved in drug development, who said it was exaggerated. The centre, based in Boston, Massachusetts, receives some funding from the pharmaceutical industry, but says that it maintains its independence. Its estimate drew on data from 10 firms on 106 randomly selected drugs that entered human testing from 1995 to 2007.

PEOPLE

New AAAS head

Rush Holt, the New Jersey physicist who is retiring from the US House of Representatives, will be the next chief executive of the American Association for the Advancement of Science (AAAS) in Washington DC. Holt replaces Alan Leshner, who since 2001 has led the organization that publishes the journal *Science*. The AAAS announced the appointment on 18 November. During his time in Congress, Holt pushed for increased science funding. Earlier in his career, he helped to lead the Princeton Plasma Physics Laboratory in New Jersey.

ITER leader

French nuclear official Bernard Bigot was nominated on 20 November as the next director-general of ITER,

the multibillion-euro project to build the world's largest nuclear-fusion reactor. Bigot (**pictured**) is currently chair of the CEA, the French Alternative Energies and Atomic Energy Commission, and will succeed Japan's Osamu Motojima. Bigot says that he plans to reform the project's management and governance, which has been blamed for budget overruns and construction delays. See go.nature.com/xct961 for more.

Jail for physicist

Turkish astrophysicist Esat Rennan Pekünlü at Ege University in Izmir begins serving a 25-month jail sentence this week, after being convicted of preventing female students who wear headscarves from attending class. In 2013, he lost his appeal to have the sentence overturned by the Constitutional Court (see go.nature.com/tm8wus). The US-based International Human Rights Network of Academies and Scholarly Societies is looking into the case, and last month Pekünlü filed an appeal to the European Court of Human Rights.

Research fraud

Igor Dzhura, a former senior research associate in biomedical engineering at Vanderbilt University in Nashville, Tennessee, falsified results in at least 69 images across 7 publications and 3 grant applications, reported the US Office of Research Integrity (ORI) on 20 November. Dzhura has also been fired by the drug firm Novartis, where he worked after leaving Vanderbilt. Investigations by the ORI and the university found that Dzhura also duplicated and renamed computer files to give the appearance of experiments he had never conducted. As part of a three-year agreement, Dzhura will retract or correct several journal articles, including a 2000 paper in *Nature Cell Biology*.

RESEARCH

Publishing pressure

The Bill & Melinda Gates Foundation announced on 20 November the world's strongest policy supporting open-access research. From January 2015, researchers funded by the charity, based in Seattle, Washington, must make their resulting papers and underlying data open access immediately on publication — and must make them available for commercial reuse. A 12-month delay on openness is allowed, if needed, until 2017. But the demand to allow commercial reuse directly conflicts with the policies of many journals, including *Nature* and *Science*. See go.nature.com/tdumvy for more.

LHC data shared

Collision data from the Large Hadron Collider (LHC) are being made freely available for the first time by CERN, Europe's particle-physics lab near Geneva, Switzerland. The Open Data Portal, launched on 20 November, is part of CERN's push to preserve its data by encouraging researchers, citizen scientists and students to mine them (see *Nature* **503,** 447; 2013). Some of the first results shared come from the LHC's Compact Muon Solenoid collaboration, which has committed to releasing data three years after collection. Selected data sets prepared for educational purposes will also be made available.

COMING UP

30 NOVEMBER
Hayabusa 2, the world's second mission to retrieve asteroid samples, launches from the Tanegashima Space Center in Japan.

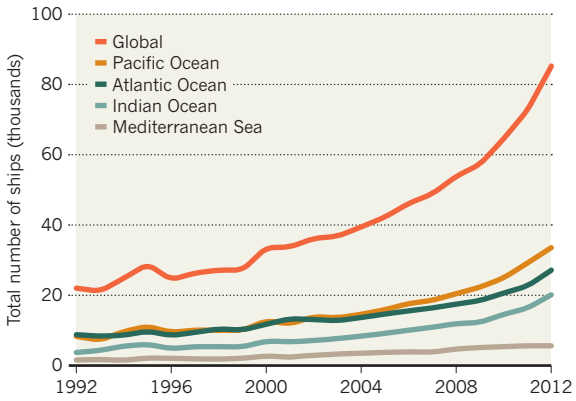
1–12 DECEMBER
Lima hosts international climate negotiations at the 20th Conference of the Parties to the United Nations Framework Convention on Climate Change.
go.nature.com/p6lv6w

TREND WATCH

Jean Tournadre, a geophysicist at France's national marine research agency IFREMER, says that he has made the most accurate determination yet of the rise in global maritime traffic — and in the accompanying pollution. Data from space-borne altimeters, which bounce radar off oceans to measure altitude, allowed him to track ships (J. Tournadre *Geophys. Res. Lett.* http://doi.org/xb5; 2014), and showed a fourfold increase in traffic over 20 years, particularly in the Arabian Sea and Bay of Bengal.

THE GROWTH IN SHIP TRAFFIC

Satellite data show global ship traffic quadrupled between 1992 and 2012, with proportional growth greatest in the Indian Ocean.



NATURE.COM
For daily news updates see:
www.nature.com/news

NEWS IN FOCUS

GREG BAKER/POOL/AP

US President Barack Obama and Chinese President Xi Jinping celebrate their joint commitment to limit carbon emissions, on 12 November.

EMISSIONS

US–China climate deal raises hopes for Lima talks

But challenges remain for United Nations meeting in run-up to a new 2015 emissions treaty.

BY JEFF TOLLEFSON

A sudden climate truce between China and the United States has renewed hopes that a two-decade stand-off between developed and developing nations over addressing climate change may at last be coming to an end. The first test will come as international negotiations resume on 1 December at the conference of the United Nations Framework Convention on Climate Change (UNFCCC) in Lima. Negotiators expect to lay the groundwork there for next year's summit in Paris, where countries are scheduled to sign a treaty that would probably take effect after 2020.

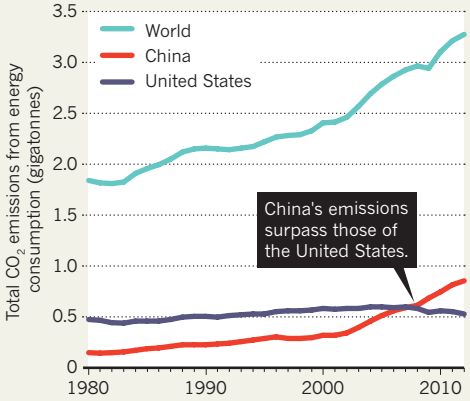
In the deal with China, US President Barack Obama committed the United States to reducing its emissions to 26–28% below 2005 levels by 2025. Chinese President Xi Jinping pledged that his country's emissions would peak around 2030, although he did not specify an exact level (see 'Carbon budget').

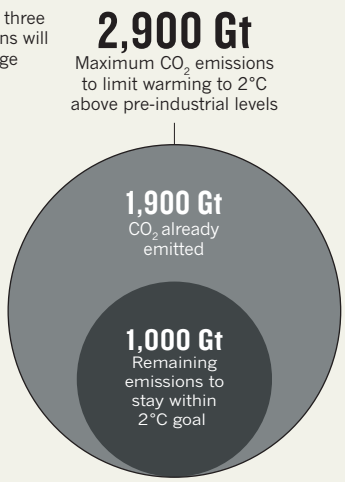
"It's hard to imagine a more important signal to get the ball rolling," says Elliot Diringer, executive vice-president of the Center for Climate and Energy Solutions, a think tank in Arlington, Virginia.

Much of the focus in Lima will be on how to translate a patchwork of emissions-reduction pledges such as the US–China deal into a fully fledged international agreement. The 2009 Copenhagen Accord brought developed and developing countries under one umbrella for the first time, but critics say that it resulted in little more than a list of promises. The question heading into the Paris summit is how to lock those national pledges into a more robust framework that includes formal procedures for verifying that countries meet their ▶

CARBON BUDGET

Global carbon dioxide emissions have risen steadily for more than three decades. With 1,900 gigatonnes (Gt) of CO_2 released already, nations will have to limit future emissions to 1,000 Gt to keep the global average temperature rise to within 2°C of pre-industrial levels.





SOURCE: UNEP

commitments, intervening when they do not and then revisiting the pledges to ratchet down emissions over time.

Saleemul Huq, director of the International Centre for Climate Change and Development in Dhaka, says that the US–China agreement marks a shift in climate negotiations. Battles may arise over the ultimate treaty's structure, further emissions reductions and financial aid promised by developed nations. But most countries, rich and poor alike, are now aligned and committed to addressing climate change. "I think momentum is the key, rather than an agreement that solves everything on paper," Huq says.

He maintains that the political challenge is quite simple: move investments away from fossil fuels and into clean energy. "If we can make that switch," he says, "then we have more or less won the game."

Nations have committed to limiting the global average temperature increase to 2 °C above pre-industrial levels, but so far there is little evidence that their governments will follow through. Scientists have calculated an overall limit on how much more carbon dioxide can be emitted to stay within that range — about a trillion tonnes — and mapped out a range of emissions-reduction pathways that are required to keep within that budget. On the basis of commitments in place before the US–China deal, the UN Environment Programme projects that emissions will exceed the level consistent with the 2 °C mark by 40% in 2030.

The US–China deal will not close this gap, but it is nonetheless a step in the right direction, says Bill Hare, managing director of Climate Analytics, a think tank in Berlin that analyses national climate policies. "They are definitely within shouting distance of the 2 °C pathway," he says, "and it's going to put pressure on many others to enter the game."

The European Union has already said that it will reduce its emissions to 40% below 1990 levels by 2030, and further pledges are expected in the first quarter of 2015. Whereas the 1997 Kyoto Protocol bound only developed countries and covered around 40% of global emissions, experts expect that the current round of commitments, scheduled to be submitted to the UNFCCC by March, includes all major countries and is likely to cover up to 80% of global emissions.

"It's a completely different ball game," says Valli Moosa, a former environment minister for South Africa and current chairman of the conservation organization WWF South Africa. He is currently co-chairing an independent dialogue involving climate officials from more than 20 leading countries at the Center for Climate and Energy Solutions — and on the basis of those discussions, he thinks that the road to Paris looks promising.

"Everything is aligned for success," he says. "Having said that, it's anybody's guess." ■

MARINE SCIENCE

Ocean observatory project hits rough water

Problems with data management challenge US effort to monitor seas in real time.

BY ALEXANDRA WITZE

From the waters off western North America to the seas surrounding Greenland, US oceanographers have nearly finished deploying hundreds of sensors, moorings, gliders and other equipment that make up an ambitious US$386-million effort to establish the world's biggest interactive portal to the oceans. Through the Ocean Observatories Initiative (OOI), set for completion by May 2015, anyone could watch the climate changing in the North Pacific or an underwater volcano erupting in real time.

But even as the final instruments splash into the sea, the project has hit a snag. After spending $37 million to develop state-of-the-art software to manage live data, the OOI has terminated that contract — shifting responsibility for 'cyberinfrastructure' from the University of California, San Diego (UCSD), to a group based at Rutgers University in New Brunswick, New Jersey.

The shift is meant to help the OOI finish construction on time and on budget by next spring, and is likely to delay the start of data streaming by only a few months. Still, it is making some oceanographers anxious, given the project's broad scope. Data from its observatories are likely to dominate US oceanography for the next 25 years — the OOI's planned lifetime.

The project, funded by the US National Science Foundation, is intended to be for the oceans what earthquake and volcano monitors are for the land: a way to allow real-time glimpses into the currents, marine life and chemistry in the deep sea. It will scrutinize a few select patches of ocean in exhaustive detail, using instruments attached to cables and on free-sailing gliders to measure temperature, salinity, chemistry and other key oceanographic parameters. The network's scattered locations include a cabled sea-floor observatory off the coast of Oregon; sets of instrumented moorings off the US east and west coasts; and

CANADIAN SCIENTIFIC SUBMERSIBLE FACILITY

Installation of the Ocean Observatories Initiative network is set to finish in March 2015.

four high-latitude sites (see 'Wired waters').

As *Nature* went to press, a few spreadsheets' worth of data gathered over the summer from the coastal sites was all that was publicly available. Live data from the entire system will start rolling out slowly over the next couple of months, says Tim Cowles, who oversees project construction for the Consortium for Ocean Leadership in Washington DC, which is overseeing the OOI.

Originally, the data were to flow into a sophisticated command-and-control system where, for instance, oceanographers could order a glider to quickly change its course. John Orcutt, head of the UCSD project, says that his team had built this command interface and that it was delivering data reliably.

But Cowles says that the project had been slipping further and further behind schedule, and so the decision was made to terminate the UCSD contract. In the early morning of 1 November, Orcutt's team shut down its connections with OOI equipment and began packing up computer servers to ship to Rutgers.

Cowles acknowledges that the Rutgers software is likely to be more basic than the UCSD system; it might, for instance, have less ability to control equipment in the water. "Somebody's going to say 'I really wanted a Ferrari and you only delivered a BMW'," he says. "But the key user functionality will be there."

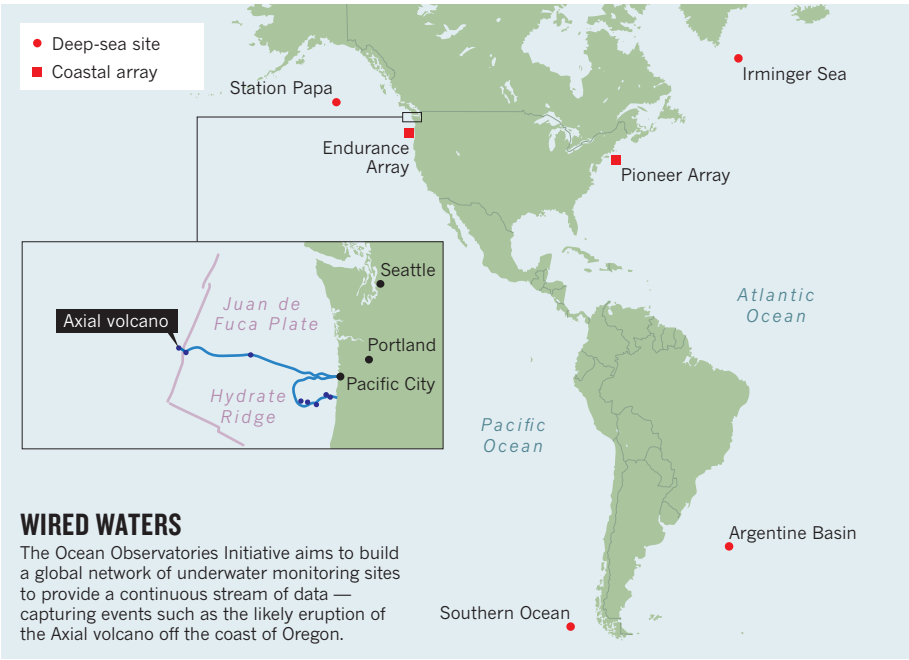
Finishing the data-management system could be a race against time, at least for a key OOI site off the Oregon coast. There, the underwater volcano Axial is building up to an eruption. This summer, in a marathon effort, a team led by scientists from the University of Washington finished installing all the instruments at Axial — including seismometers, web cameras, and others — for the sea-floor cabled observatory. "It's in the water, and it works," says John Delaney, an oceanographer at the University of Washington in Seattle who designed the observatory. "I'm very, very excited about that." But data from the OOI cable are not yet available to the broader scientific community.

The timing is urgent: Axial last erupted in 2011, and magma is now refilling the sea floor beneath the volcano so quickly that some geologists predict that it will erupt within a year. As they wait for the OOI software, Delaney and others are working to arrange access through a data centre at the Incorporated Research Institutions for Seismology in Washington DC. "I want the data in the hands of the community," says Delaney. "That's what it was all about to begin with."

In addition to the unfinished data pipeline, two major parts of the OOI have yet to go into the water. In February and March 2015, oceanographers aboard the research vessel *Atlantis* will deploy moored sensor arrays off the coasts of Chile and Argentina — the network's final sites.

The OOI will be formally commissioned in mid-May, which is the ultimate deadline for all the data to be publicly available. Science workshops on using the data will take place starting early next year, Cowles wrote in a 24 November web update.

"If there are expectations that were developed six years ago by members of the community, maybe that wish won't necessarily come true based on their earliest dream," says Cowles. "Nonetheless, the built reality of the OOI will be remarkable, and unlike any functionality we in ocean science have had before." ■



WIRED WATERS

The Ocean Observatories Initiative aims to build a global network of underwater monitoring sites to provide a continuous stream of data — capturing events such as the likely eruption of the Axial volcano off the coast of Oregon.

UPMC

Patients can now guide robotic limbs using devices implanted in their brains.

NEUROSCIENCE

Regulators assess brain–linked devices

Food and Drug Administration homes in on neural implants.

BY SARA REARDON

For the first time since accidents severed the neural connection between their brains and limbs, a small number of patients are reaching out and feeling the world with prosthetic devices wired directly to their brains. Earlier this month, scientists at the California Institute of Technology (Caltech) in Pasadena implanted a person's brain with electrode arrays that read neural activity to control a robotic arm and stimulate the brain to deliver a sensation of what the arm touched. And since 2011, a team at the University of Pittsburgh in Pennsylvania has been working with a small number of people who control prostheses through neural implants. "It's moving quick at the moment," says Christian Klaes, a neuroscientist on the Caltech effort. "The race has started."

The advances are also starting to attract serious attention from the US Food and Drug Administration (FDA), which is wrestling with how best to regulate such brain–computer interfaces to ensure that they are safe. On 21 November, the agency held a meeting at its White Oak campus in Silver Spring, Maryland, to get the process started. The meeting was well-timed: in May, the FDA approved a robotic arm that can be controlled with brain implants. And the US Defense Advanced Research Projects Agency is funding the development of prosthetic devices that read brainwaves, as well as implants that electrically stimulate organs to perform functions such as insulin production.

At the meeting, researchers discussed their progress and challenges, and FDA experts talked about the regulatory steps required to bring a device to market. The riskiest devices are implants that require brain surgery; these have been inserted in only a few people under controlled laboratory settings. Less-invasive devices, such as electrodes that sit on the head, or sensors that pick up electrical activity from muscles, are further along the developmental path. FDA scientists presented some of their own research on safety questions, such as how long electrodes can safely stay in the brain and what happens when they malfunction.

Researchers have welcomed the agency's willingness to consider broader use of the devices, especially given that experiments in small laboratory settings seem promising — and are important for understanding the challenges involved. At the Society for Neuroscience meeting in Washington DC this month, Klaes and his colleagues presented the first year of data from an electrode array implanted in a person's posterior parietal cortex, a brain region that controls the intent to perform an action rather than its actual execution. Their patient has learnt to control avatars in video games and uses a brain-controlled arm to play the classic hand game rock–paper–scissors. Another person, who received a motor-cortex implant from the University of Pittsburgh team in 2012 to control an arm, can now perform tasks such as feeding herself[1].

But the companies that develop the devices are also slightly anxious about the FDA's interest, because the agency's requirements for safety and effectiveness set a high bar for any new device. Utah-based Blackrock Microsystems in Salt Lake City, which manufactures the electrode arrays being used by the Caltech and Pittsburgh groups, is opening a branch this week in Hanover, Germany, partly because regulations there are less stringent. "We need some clear guidelines, even at the engineering phase," says chief financial officer Marcus Gerhardt. The company sometimes hesitates over developing new devices if it thinks that the FDA will not approve them, he says.

Regulations aside, companies are still weighing up the advantages of certain devices, such as implants in the somatosensory cortex, which enable a patient to register sensations. Klaes says that the somatosensory implants work well in monkeys, allowing the animals to rummage through a "handbag" to find sensory targets they cannot see[2].

But the market for brain-connected prosthetics is very small: people who have lost control over much or all of their bodies are few and far between. For those who have had amputations — a much larger population — a prosthetic arm with no direct connection to the brain is much easier to control, and often sufficient.

And brain implants might not always be necessary for delivering sensory signals from a prosthesis. In October, neural engineer Dustin Tyler at Case Western Reserve University in Cleveland, Ohio, described a hand prosthesis that stimulates nerves in an arm stump at different frequencies to simulate different textures[3].

"We need some clear guidelines, even at the engineering phase."

Even if the FDA approves the devices, health insurers must be convinced that they are necessary — and agree to pay for them — before it makes business sense for firms to manufacture them. And the FDA has to be sure that the functional improvement outweighs safety concerns. "From the patient's point of view, of course, they would like to have their own arm," says Kip Ludwig, a programme director at the National Institute of Neurological Disorders and Stroke in Bethesda, Maryland. "But that's a long way away." ■

1. Collinger, J. L. *et al. Lancet* **381,** 557–564 (2013).
2. Klaes, C. *et al. J. Neural Eng.* **11,** 056024 (2014).
3. Tan, D. W. *et al. Sci. Transl. Med.* **6,** 257ra138 (2014).

MEDICINE

Clinical-trial rules to improve access to results

Agencies propose expanded reporting of drug-test data.

BY SARA REARDON

The US website ClinicalTrials.gov is the world's largest repository of clinical-trial information, containing the results of more than 179,000 studies conducted in 187 countries. Yet the database represents only a fraction of the trials that are run. Despite US laws requiring that results be posted to the site, drug companies and academic researchers have found numerous ways to withhold data that show that a drug did not work or had serious side effects.

Now regulators are trying to close some of these loopholes. On 19 November, the US National Institutes of Health (NIH) and the Food and Drug Administration (FDA) proposed regulations that would tighten reporting requirements and expand financial penalties for violating them. "When a lot of dollars and time and volunteers are potentially putting themselves in a risk situation, we need to be sure the results of that are finding their way into view of the public," NIH director Francis Collins said at a press conference to announce the regulations.

Transparency is necessary, he said, to ensure that study volunteers can find out what is known about a treatment and its side effects when deciding whether to participate in a trial. Researchers benefit as well, because complete reporting of data allows them to build on the results and avoid repeating failed experiments.

But although some bioethicists praise the initiative, many worry that the new rules will not solve the underlying problem. "In a lot of ways, coming up with new regulations is really the easy part," says Christopher Jones, a physician at Rowan University in Camden, New Jersey. "The more difficult and more important phase is going to be making sure that the new regulations are enforced fairly and transparently."

The proposed rules would broaden a 2007 law known as the FDA Amendments Act (FDAAA), which requires researchers to post the results of studies involving FDA-approved drugs on ClinicalTrials.gov within 30 days of approval. But such results may never be posted if a drug is never approved. One new proposal from the government would require companies to post results for all drugs or therapies submitted for FDA approval.

> ***"This shows a much broader understanding of what a clinical trial is."***

"It's a great step," says Jones. "Industry in general is pretty good at following regulations as long as regulations are clearly stated and enforced."

But Jennifer Miller, a bioethicist at Duke University in Durham, North Carolina, says that the majority of trial sponsors are not following the current law. Data she has compiled comparing the number of trials registered with the FDA to those reported on ClinicalTrials.gov suggest that many are not reported, even for drugs that have been approved. The FDA can levy a fine of US$10,000 a day for noncompliance, but has never done so. "If you were going to expand or enhance FDAAA, you would think there would be considerations around monitoring and enforcement of the existing law," Miller says.

PHASE TWO

A second proposal released by the government this week would go a step further by requiring federally funded researchers to post the results of phase I clinical trials, which focus on the safety of a drug or device rather than its efficacy. Researchers would also have to register and report the results of studies that evaluate surgical techniques or non-medical interventions, such as public-policy changes intended to curb smoking. The NIH says that it plans to enforce this rule among its own researchers, and could withdraw funding from external institutions that do not comply. The agency would first try to resolve the problem with an institution before taking this step, says Sally Rockey, NIH deputy director for extramural research.

"I think there's a lot of good here," says Kay Dickersin, director of the Center for Clinical Trials at Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland. "This shows a much broader understanding of what a clinical trial is." Still, she and Jones say that they wish that studies sponsored by industry and private sources were also required to report results of phase I trials. Pharmaceutical companies often guard these results, because they can contain proprietary information. Jones says that even if a drug never makes it to market, revealing the results of early safety tests could save time and money for researchers trying to develop similar therapies in the future.

Dickersin worries that it will still be hard to find information about unwanted side effects and other adverse events under the proposed regulations. Trial sponsors are required only to report summary data about volunteers' reactions to a drug, not each person's results. But data on individuals allows outside researchers to do independent analyses that may yield different conclusions (M. A. Rodgers *et al. Br. Med. J.* **346,** f3981; 2013).

This may change, says Kathy Hudson, deputy director for science, outreach and policy at the NIH. The US Institute of Medicine is working on a report weighing the downsides of summaries against other concerns such as volunteers' privacy. That report is expected to be released in early 2015. For now, the public has 90 days to comment on the proposed clinical-trial regulations before they become law. ■

MORE ONLINE

INTERVIEW

MORTON BEEBE/CORBIS

Forty years of Lucy, an ancient ancestor who people can relate to go.nature.com/ojipbl

MORE NEWS

- Robot reveals surprisingly thick Antarctic sea ice go.nature.com/rtp1qu
- Pain and itch grown in a dish go.nature.com/jkrpfv
- Crowd-funded Moon mission is serious about science go.nature.com/nc8eli

NATURE PODCAST

Do-it-yourself peer review, powerless air-con, and sexology on display nature.com/nature/podcast

PUBLISHING

Confusion over open-data rules

The Public Library of Science's pioneering open-data mandate has prompted scientists to share more data online, but not everyone is complying with the regulation.

BY RICHARD VAN NOORDEN

The mantra of the nascent open-data movement — that scientists should share online all data underlying their findings — sounds simple. But it can be tough to achieve in practice. An informal audit of one of the movement's biggest proponents, the Public Library of Science (PLOS), shows that not everyone is complying with the publisher's pioneering open-data mandate, and hints at the challenges that journals can face in enforcing open-data goals.

The idea that the progress of research will be accelerated if others can easily and freely build on data sets is gaining currency. Last week the Bill & Melinda Gates Foundation in Seattle, Washington, announced that it would demand open data of the researchers it funds.

But whereas some research communities, such as geneticists and crystallographers, have long-established norms of open data, most funders and publishers (including *Nature*), mindful of researcher autonomy, merely exhort scientists to make their data open. Many surveys have found that scientists are worried about being scooped on future projects, or argue that they have signed agreements not to share their data.

So it was a step-change when in March PLOS made it a requirement that authors who publish in its journals share online all the data necessary to reproduce their studies. It was not the first publisher to convert encouragement into a mandate, but it was the largest.

The new policy piqued the curiosity of Tim Vines, managing editor of *Molecular Ecology*, one of the few journals apart from the PLOS family with an open-data mandate. Vines and his colleagues published a survey on the effectiveness of different open-data mandates in early 2013 (T. H. Vines *et al. FASEB J.* **27,** 1304–1308; 2013), focusing on a subset of evolutionary biology papers that all used a free software package called STRUCTURE to map the genetic structure of populations on the basis of DNA profiling of individuals.

"A complete culture shift will be further down the line."

That study included 51 *PLoS ONE* papers, and found that just 6 of them had shared the data that went into the STRUCTURE study. In a new analysis, Vines found 20 papers that mentioned STRUCTURE and had been published since March 2014 — including one that tracked different varieties of cotton plants in the Caribbean, and another that compared different populations of a particular sparrow across the southern United States. Eight of the new studies (40%) had shared the genotype data — meaning that a reader would be able to repeat their analysis. The remaining 60% of papers had not made their data available, even though each stated that "all data underlying the findings are fully available without restriction", in accordance with PLOS's policy (see 'Free the data').

Because the open-data mandate is new, and *PLoS ONE* publishes so many papers, some manuscripts do get published without all their data being made available, says the journal's editorial director Damian Pattinson. *PLoS ONE* does perform internal checks for some data types, but in this case it would have been the job of external peer reviewers to check whether appropriate underlying data were available, he says. And although *PLoS ONE* was grateful for their help, "there is a learning curve here for all involved to understand what the data-sharing standards are for all disciplines and data types". He adds that once someone complains, the journal has a system for investigating papers that do not comply with its open-data requirement.

The research teams in question, nine of whom responded to the *Nature* news team's request for an explanation, provide an insight into why data sharing does not always happen in the first place (full data are available at go. nature.com/8ggob6). Some had forgotten to upload their data and promptly rectified the fault. Four teams said that the journal's editors and referees had never asked them to share the underlying genetics data, suggesting confusion over what the policy means.

Others aired wider — and common — objections to online data sharing. Even though they had chosen to publish with PLOS, some authors said that they wanted to hold back their data for future studies, or did not want to share the raw data unless they knew future users' intent.

Steve Simpson at the University of Exeter, UK, who reported work on Omani clownfish, said he was happy to share raw data privately with potential collaborators, but not to upload results from 400 individual fish that had taken great effort to collect. "The study is described so that it could be replicated by another expeditionary team who were willing to dive across Oman collecting rare fish under several hard-earned licences," he wrote.

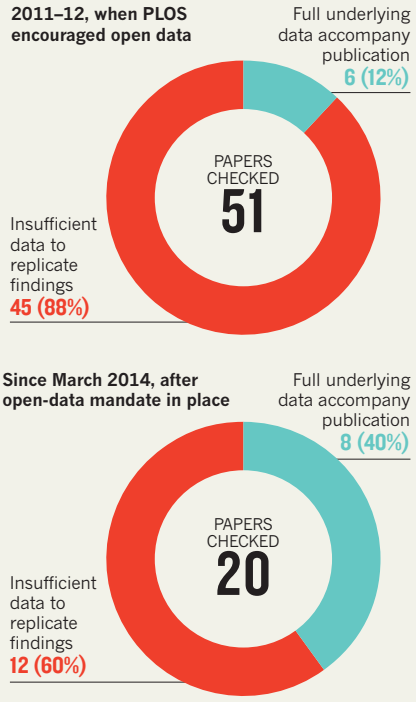
"There is a lot of inconsistency among fields as to what data are shared," said another researcher, Sabrina Taylor at Louisiana State University in Baton Rouge, who conducted the sparrow study. She had not uploaded her genetics data because, she said, she was not aware of a public data repository for it.

Vines is optimistic about the prospects for open data. The PLOS mandate means the situation is "already better than it was", he says. "At its core, the problem is author education", he adds. Even at *Molecular Ecology*, which has been enforcing an open-data policy since 2011, "we still have to ask for additional data sets for about half of papers at the acceptance stage".

It will take time to make PLOS's policy clear and easy to comply with across all scientific fields, says Pattinson. "A complete culture shift will be further down the line." ■

FREE THE DATA

In 2014, open-access publisher PLOS introduced a requirement that authors who publish in its journals make their underlying data freely available online. An informal audit of one type of population genetics study in one journal, *PLoS ONE*, shows that not everyone is complying — but the mandate is still a boon for the open-data movement.



SOURCE: TIM VINES

JOEL SARTORE/NATL GEOGRAPHIC/GETTY

Tortoises at the Charles Darwin Research Station on Santa Cruz in the Galapagos Islands.

ECOLOGY

Key Galapagos research station in trouble

Local government's closure of gift shop could doom Charles Darwin Foundation.

BY ALESZU BAJAK

For more than half a century, the Charles Darwin Foundation (CDF) has supported a thriving research station in Ecuador's Galapagos Islands. Scientists at the station have helped to bring the iconic Galapagos tortoise back from the brink of extinction and to eradicate invasive goats from Isabela, the largest island in the Galapagos archipelago.

But that long legacy is being threatened by a spat with the local government, which could force the Charles Darwin Research Station to close. In July, officials on Santa Cruz island ordered the CDF to shut its lucrative gift shop in the town of Puerto Ayora, citing complaints from restaurateurs and shop owners who said that the store was siphoning away their business. That has deprived the foundation of at least US$8,000 per week in income; total losses could reach $200,000 if the shop remains closed for the rest of the year, the foundation says.

"The closure of the store basically ruined our 2014 budget," says CDF president Dennis Geist, a volcanologist who has studied Galapagos sites for 30 years. "We have no endowment. We don't even have any reserve funds. The closing of the Darwin station is a very realistic possibility right now."

On 24 November, the CDF's governing body met in Quito, Ecuador. Its voting members, who include employees of the Ecuadorian federal government, agreed to form a working group "to strategically secure the operation of the research station".

But the financial troubles are already affecting operations at the station, which employs around 65 people and works with more than 100 international scientific collaborators. Although the gift shop provides just 10% of the foundation's revenue, its closure has had cascading effects, says CDF executive director Swen Lorenz. "We have already lost a significant donation from someone who said that if the government of Ecuador doesn't support us having a souvenir shop, then he won't support us with a donation," he says. "We're two and a half months late with salary, projects haven't been running, and we've had one staff member leave."

Alex Hearn, director of conservation science at the Turtle Island Restoration Network, an environmental advocacy group based in Olema, California, says that the closure of Darwin station would be a major blow. Nearly every scientist who has worked in the Galapagos has dealt either directly or indirectly with the CDF, says Hearn, who coordinated fisheries research at the station from 2002 to 2008.

He still works closely with scientists there on fisheries and shark research. "I don't have to jump on a plane every time I need some data," he says. "I know the research can be done, and done well." ■

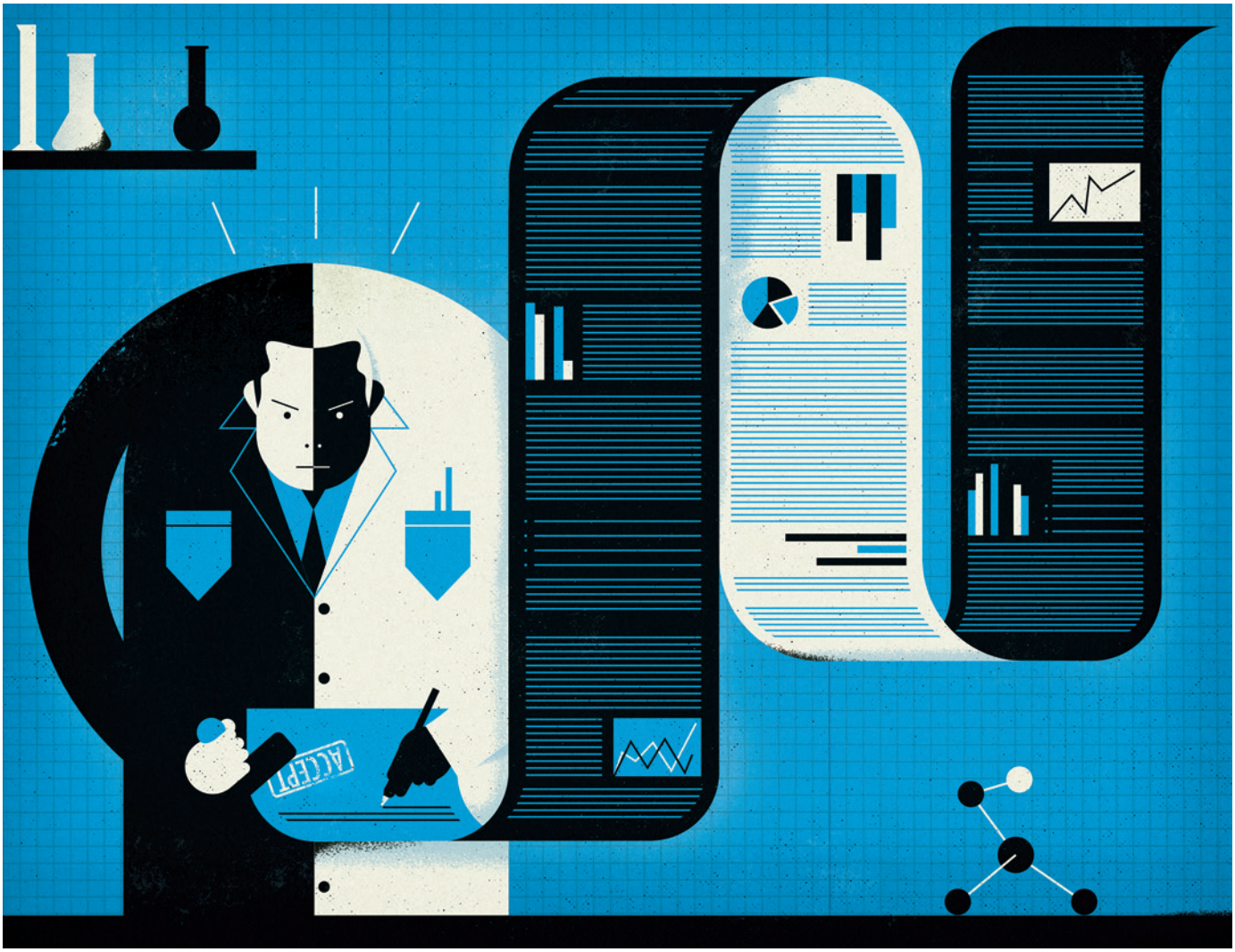


ILLUSTRATION BY DALE EDWIN MURRAY

THE PEER-REVIEW SCAM

When a handful of authors were caught reviewing their own papers, it exposed weaknesses in modern publishing systems. Editors are trying to plug the holes.

BY CAT FERGUSON, ADAM MARCUS AND IVAN ORANSKY

Most journal editors know how much effort it takes to persuade busy researchers to review a paper. That is why the editor of *The Journal of Enzyme Inhibition and Medicinal Chemistry* was puzzled by the reviews for manuscripts by one author — Hyung-In Moon, a medicinal-plant researcher then at Dongguk University in Gyeongju, South Korea.

The reviews themselves were not remarkable: mostly favourable, with some suggestions about how to improve the papers. What was unusual was how quickly they were completed — often within 24 hours. The turnaround was a little too fast, and Claudiu Supuran, the journal's editor-in-chief, started to become suspicious.

In 2012, he confronted Moon, who readily admitted that the reviews had come in so quickly because he had written many of them himself. The deception had not been hard to set up. Supuran's journal and several others published by Informa Healthcare in London

invite authors to suggest potential reviewers for their papers. So Moon provided names, sometimes of real scientists and sometimes pseudonyms, often with bogus e-mail addresses that would go directly to him or his colleagues. His confession led to the retraction of 28 papers by several Informa journals, and the resignation of an editor.

Moon's was not an isolated case. In the past 2 years, journals have been forced to retract more than 110 papers in at least 6 instances of peer-review rigging. What all these cases had in common was that researchers exploited vulnerabilities in the publishers' computerized systems to dupe editors into accepting manuscripts, often by doing their own reviews. The cases involved publishing behemoths Elsevier, Springer, Taylor & Francis, SAGE and Wiley, as well as Informa, and they exploited security flaws that — in at least one of the systems — could make researchers vulnerable to even more serious identity theft. "For a piece of software that's used by hundreds of thousands of academics worldwide, it really is appalling," says Mark Dingemanse, a linguist at the Max Planck Institute for Psycholinguistics in Nijmegen, the Netherlands, who has used some of these programs to publish and review papers.

But even the most secure software could be compromised. That is why some observers argue for changes to the way that editors assign papers to reviewers, particularly to end the use of reviewers suggested by a manuscript's authors. Even Moon, who accepts the sole blame for nominating himself and his friends to review his papers, argues that editors should police the system against people like him. "Of course authors will ask for their friends," he said in August 2012, "but editors are supposed to check they are not from the same institution or co-authors on previous papers."

PEER-REVIEW RING

Moon's case is by no means the most spectacular instance of peer-review rigging in recent years. That honour goes to a case that came to light in May 2013, when Ali Nayfeh, then editor-in-chief of the *Journal of Vibration and Control*, received some troubling news. An author who had submitted a paper to the journal told Nayfeh that he had received e-mails about it from two people claiming to be reviewers. Reviewers do not normally have direct contact with authors, and — strangely — the e-mails came from generic-looking Gmail accounts rather than from the professional institutional accounts that many academics use (see 'Red flags in review').

Nayfeh alerted SAGE, the company in Thousand Oaks, California, that publishes the journal. The editors there e-mailed both the Gmail addresses provided by the tipster, and the institutional addresses of the authors whose names had been used, asking for proof of identity and a list of their publications. One scientist responded — to say that not only had he not sent the e-mail, but he did not even work in the field.

This sparked a 14-month investigation that came to involve about 20 people from SAGE's editorial, legal and production departments. It showed that the Gmail addresses were each linked to accounts with Thomson Reuters' ScholarOne, a publication-management system used by SAGE and several other publishers, including Informa. Editors were able to track every paper that the person or people behind these accounts had allegedly written or reviewed, says SAGE spokesperson Camille Gamboa. They also checked the wording of reviews, the details of author-nominated reviewers, reference lists and the turnaround time for reviews (in some cases, only a few minutes). This helped the investigators to ferret out further suspicious-looking accounts; they eventually found 130.

As they worked through the list, SAGE investigators realized that authors were both reviewing and citing each other at an anomalous rate. Eventually, 60 articles were found to have evidence of peer-review tampering, involvement in the citation ring or both. "Due to the serious nature of the findings, we wanted to ensure we had researched all avenues as carefully as possible before contacting any of the authors and reviewers," says Gamboa.

When the dust had settled, it turned out that there was one author in the centre of the ring: Peter Chen, an engineer then at the National Pingtung University of Education (NPUE) in Taiwan, who was a co-author on practically all of the papers in question. After "a series of unsatisfactory responses" from Chen, says Gamboa, SAGE contacted the NPUE, which joined the investigation into Chen's work. Chen resigned from his post in February 2014.

In May, Nayfeh resigned over the scandal at his journal, and SAGE contacted the authors of all 60 affected articles to let them know that the papers would be retracted. Chen could not be reached for comment for this story, but Taiwan's state-run news agency said in July that he had issued a statement taking sole responsibility for the peer-review and citation ring, and admitting to the "indiscreet practice" of adding Taiwan's education minister as a co-author on five of the papers without his knowledge. That minister, Chiang Wei-ling, denies any involvement, but nevertheless resigned "to uphold his own reputation and avoid unnecessary disturbance of the work of the education ministry", according to a public statement.

The collateral damage did not stop there. A couple of authors have asked SAGE to reconsider and reinstate their papers, Gamboa says, but the publisher's decision is final — even if the authors in question knew nothing of Chen or the peer-review ring.

PASSWORD LOOPHOLE

Moon and Chen both exploited a feature of ScholarOne's automated processes. When a reviewer is invited to read a paper, he or she is sent an e-mail with login information. If that communication goes to a fake e-mail account, the recipient can sign into the system under whatever name was initially submitted, with no additional identity verification. Jasper Simons, vice-president of product and market strategy for Thomson Reuters in Charlottesville, Virginia, says that ScholarOne is a respected peer-review system and that it is the responsibility of journals and their editorial teams to invite properly qualified reviewers for their papers.

Nature Publishing Group (NPG) owns a few journals that use ScholarOne, but *Nature* itself and *Nature*-branded journals use different software, developed by eJournalPress of Rockville, Maryland. Véronique Kiermer, *Nature*'s executive editor and director of author and reviewer services for NPG in New York City, says that NPG does not seem to have been the victim of any such peer-review-rigging schemes.

But ScholarOne is not the only publishing system with vulnerabilities. Editorial Manager, built by Aries Systems in North Andover, Massachusetts, is used by many societies and publishers, including Springer and PLOS. The American Association for the Advancement of Science in Washington DC uses a system developed in-house for its journals *Science*, *Science Translational Medicine* and *Science Signaling*, but its open-access offering, *Science*

RED FLAGS IN REVIEW

Signs that an author might be trying to game the system

A handful of researchers have exploited loopholes in peer-review systems to ensure that they review their own papers. Here are a few signs that should raise suspicions.

- The author asks to exclude some reviewers, then provides a list of almost every scientist in the field.
- The author recommends reviewers who are strangely difficult to find online.
- The author provides Gmail, Yahoo or other free e-mail addresses to contact suggested reviewers, rather than e-mail addresses from an academic institution.
- Within hours of being requested, the reviews come back. They are glowing.
- Even reviewer number three likes the paper.

Advances, uses Editorial Manager. Elsevier, based in Amsterdam, uses a branded version of the same product, called the Elsevier Editorial System.

Editorial Manager's main issue is the way it manages passwords. When users forget their password, the system sends it to them by e-mail, in plain text. For *PLOS ONE*, it actually sends out a password, without prompting, whenever it asks a user to sign in, for example to review a new manuscript. Most modern web services, such as Google, hide passwords under layers of encryption to prevent them from being intercepted. That is why they require users to reset a password if they forget it, often coupled with checking identity in other ways.

Security loopholes can do more than compromise peer review. Because people often use the same or similar passwords for many of their online activities — including banking and shopping — e-mailing out the password presents an opportunity for hackers to do more than damage the research record. Dingemanse, who has published in a number of journals that use Editorial Manager, including *PLOS ONE*, says: "It's quite amazing that they haven't got around to implementing a safe system." Neither Aries nor *PLOS ONE* responded to several requests for comment.

SAFETY MEASURES

Lax password protection has resulted in breaches. In 2012, the Elsevier journal *Optics & Laser Technology* retracted 11 papers after an unknown party gained access to an editor's account and assigned papers to fake reviewer accounts. The authors of the retracted papers were not implicated in the hack, and were offered the chance to resubmit.

Elsevier has since taken steps to prevent reviewer fraud, including implementing a pilot programme to consolidate accounts across 100 of its journals. The rationale is that reducing the number of accounts in its system might help to reveal those that are fraudulent, says Tom Reller, a spokesperson for Elsevier. If it is successful, consolidation will roll out to all journals in early 2015. Furthermore, passwords are no longer included in most e-mails from the editorial system. And to verify reviewers' identities, the system now integrates the Open Researcher and Contributor ID (ORCID) at various points. ORCID identifiers, unique numbers assigned to individual researchers, are designed to track researchers through all of their publications, even if they move institutions.

ScholarOne also allows ORCID integration, but it is up to each journal to decide how to use it. Gamboa says that not enough scientists have adopted the system to make it possible to require an ORCID for each reviewer. And there is another problem: "Unfortunately, like any online verification system, ORCID is also open to the risk of unethical manipulation," says Gamboa — for example, through hacking.

That is a common refrain. "As you make the system more technical and more automated, there are more ways to game it," says Bruce Schneier, a computer-security expert at Harvard Law School's Berkman Center for Internet and Society in Cambridge, Massachusetts. "There are almost never technical solutions to social problems."

It ultimately falls to editors and publishers to be on the alert, particularly when contacting potential reviewers. Carefully checking e-mail addresses is one way to ferret out fakes: a non-institutional e-mail address such as a free account from Gmail is a red flag, say sources. But at the same time, it could also be a perfectly legitimate address.

> "As you make the system more technical and more automated, there are more ways to game it."

Jigisha Patel, associate editorial director of BioMed Central in London, says that it is definitely possible to catch cheaters by being on the alert for dubious e-mail addresses. "We've had some cases where we've caught them tweaking the e-mail addresses to try to steal someone's identity," she says. But such screening is imperfect. In September, the publisher retracted a paper in *BMC Systems Biology*, stating that it believed that "the peer-review process was compromised and inappropriately influenced by the authors".

Some scientists and publishers say that journals should not allow authors to recommend reviewers in the first place. John Loadsman, an editor of *Anaesthesia and Intensive Care*, which is published by the Australian Society of Anaesthetists in Sydney, calls the practice "bizarre" and "completely nuts", and says that his journal does not permit it.

It is unclear exactly what proportion of journals allows the practice, but as fields become more specialized it provides an easy way for busy editors to find relevant expertise. Jennifer Nyborg, a biochemist at Colorado State University in Fort Collins, says that most of the journals to which she submits articles request at least five potential reviewers.

For most of the 60 articles retracted by SAGE, the original peer review had used only author-nominated reviewers. Despite this experience, the *Journal of Vibration and Control* still allows authors to suggest peer reviewers (and provide their contact e-mails) when they submit a manuscript — although more safeguards are now in place, says Gamboa.

The Committee on Publication Ethics (COPE), which serves as a kind of moral compass for scientific publishing (but has no authority to enforce its advice) has no guidance on the practice, but urges journals to vet reviewers adequately. Good practice is always to check the names, addresses and e-mail contacts of reviewers, says Natalie Ridgeway, operations manager for COPE in London. "Editors should never use only the preferred reviewer."

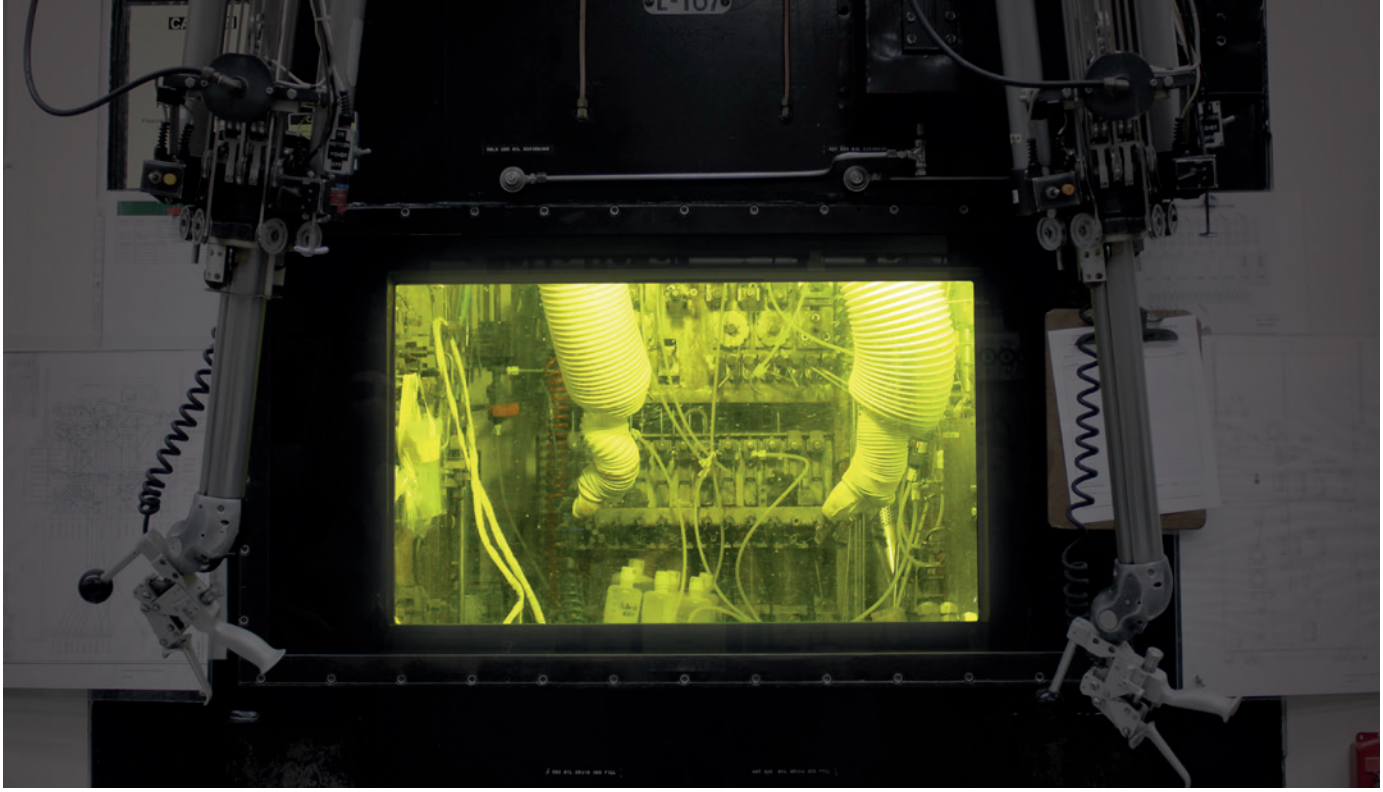
NPG journals do allow authors to suggest independent reviewers. "But these suggestions are not necessarily followed," says Kiermer. "The editors select reviewers and the selection includes checking for the absence of conflict of interests." On the flip side, authors can ask an editor to exclude reviewers who they believe to have unmanageable conflicts, such as competing research. The publisher usually honours such requests, as long as authors do not ask to exclude more than three people or labs, Kiermer says.

Sometimes, recommending reviewers can backfire. Robert Lindsay, one of two editors-in-chief of the Springer-published journal *Osteoporosis International*, says that his publication allows authors to recommend up to two reviewers — but that he often uses this information to rule those reviewers out. This is based on past experience, in which he has seen authors recommend their own contacts, or worse: "We have had family members, folks in the same department, postgraduate students being supervised by an author," he says. The journal generally uses suggested reviewers — who have passed screening — only if it runs into trouble finding other scientists to perform the task.

But screening can be difficult. Usually, editors in the United States and Europe know the scientific community in those regions well enough to catch potential conflicts of interest between authors and reviewers. But Lindsay says that Western editors can find this harder with authors from Asia — "where often none of us knows the suggested reviewers". In these cases, the journal insists on at least one independent reviewer, identified and invited by the editors.

In what Lindsay calls the worst case that he has seen, an author suggested a reviewer who shared her first name but not her surname. Some investigation revealed that the surname was the author's maiden name — she was recommending that she review her own paper. "I don't think she is going to submit anything to us again," says Lindsay. ■

Cat Ferguson, **Adam Marcus** *and* **Ivan Oransky** *are the staff writer and two co-founders, respectively, of Retraction Watch in New York City.*



ALEXANDRA WITZE

DESPERATELY SEEKING PLUTONIUM

NASA has 35 kilograms of plutonium-238 to power its deep-space missions — but that will not get it very far.

BY ALEXANDRA WITZE

Ken Wilson peers through a yellow-tinted window at the clutter of bottles and chemical equipment on the other side. He is protected from the radiation their contents are giving off by five thick panes of glass interspersed with some 400 litres of oil.

Working in such a 'hot cell' is routine for Wilson, who is one of the top nuclear technicians here at the Oak Ridge National Laboratory (ORNL) in Tennessee. Grasping the handholds of some robotic manipulator arms, he begins to move them like extensions of his own lanky frame — first picking up bottles inside the cell, then uncapping them and pouring liquids from one container to another.

Eventually, Wilson will add the residue of all this remote-controlled chemistry to a dark-brown liquid that fills two bottles sitting off in the hot cell's corner. This liquid is a concentrated solution of plutonium-238: a highly radioactive isotope that was made here at Oak Ridge, and that Wilson is now working to purify. Its ultimate destination is deep space, where heat from its decay will power NASA missions such as future Mars rovers, or spacecraft heading to the outer Solar System, where the Sun's rays can be too dim for solar panels.

NASA will be relieved to get this ^{238}Pu, because it is increasingly anxious about running out. The isotope is not found in nature, so it has to be made in nuclear reactors. But the main US supply shut down in 1988, when the Savannah River Plant near Aiken, South Carolina, run by the Department of Energy (DOE), stopped making ^{238}Pu as part of a nuclear-weapons phase-out. Four years later, the DOE began purchasing small amounts of the isotope from the Russian government, but those acquisitions have also ended.

As a result, NASA now has just 35 kilograms of plutonium product — a small supply that may not match the demand to send missions to Mars, the moons of Jupiter and beyond. And the crunch got even worse in late 2013, when budget constraints led NASA to cancel a programme to develop a radioisotope power source that would have used one-quarter of the plutonium of conventional designs (see *Nature* http://doi.org/w8m; 2013).

This is why Wilson is doing chemistry in the Oak Ridge hot cells. Last year, in a move that was unprecedented for both agencies, NASA started paying the DOE US$50 million a year to reactivate its long-stalled capability for making ^{238}Pu. That is a tall order: the DOE is now grappling with having to produce the material in facilities that were never set up for it; interviewing retired plutonium technicians for tips on how to manufacture and store the isotope; and designing machines and workflows that can accommodate more than a kilogram of plutonium per year moving through the system.

"The plutonium-production business is hard to do," says Ralph McNutt, a planetary scientist at the Johns Hopkins University Applied Physics Laboratory in Laurel, Maryland, who is participating in an internal NASA study on developing nuclear power for space missions. "Everybody took it for granted that it was out there and would always be there. Life's a little more complicated than that."

HOT ZONE

The first radioisotope power units were developed in the late 1950s and early 1960s by the US and Soviet space programmes. (The European Space Agency has never developed nuclear power sources for missions, a policy that limited the operating life of a solar-powered lander that visited a comet earlier in November.) The United States has used radioisotope power units on 27 missions, from a Navy navigation satellite launched in 1961 to the Mars Curiosity rover in 2011.

All follow the same basic idea: as the isotope decays, the radioactivity heats the junction between two metals or semiconductors (see 'Power trip'). Thanks to a phenomenon known as the thermoelectric effect, this sets up an electric current that the spacecraft can use to power its instruments, or store in a battery.

A 'hot cell' at Oak Ridge National Laboratory, where plutonium is processed.

Smaller radioisotope units can also help to keep a probe warm in the frigid environment of space.

The isotope of choice is ^{238}Pu, partly because it produces a high amount of power per gram of material, and partly because of worker safety: it emits only α-particles, which are relatively easy to shield against.

NASA's current favoured design for a nuclear power source, the Multi-Mission Radioisotope Thermoelectric Generator (MMRTG), uses 4.8 kilograms of plutonium dioxide — a chemically stable compound — to provide 2,000 watts of heat and 110 watts of electrical power at a mission's start. With a half-life of 87.7 years, ^{238}Pu can produce power for decades. But the output fades over time. Project scientists working with the Voyager 1 spacecraft, which was launched in 1977 and is now more than 19 billion kilometres from Earth, have had to turn off instruments one by one as the electricity from its power units has dwindled.

With 35 kilograms of plutonium dioxide on the shelf, NASA might seem in a good position to fuel many future nuclear-powered spacecraft. But the stockpile has aged, and less than half of it now meets NASA specifications in terms of how much heat it produces. Given the long lead time in planning planetary missions, and the challenges in maintaining the plutonium supply for missions not yet even dreamed of, the agency is less well-off than it might appear.

NASA will use about 5 kilograms as a generator on the next Mars rover, set to launch in 2020. And future missions to the outer Solar System could require multiple generators.

The new contract with the DOE will for the first time provide NASA with a steady supply of the isotope. The goal is for the DOE to produce 1.5 kilograms of plutonium dioxide a year by 2021, which translates to about 1.1 kilograms a year of ^{238}Pu. With that small influx, NASA should have enough to fuel about two missions a decade, says David Schurr, deputy director of NASA's planetary-sciences division in Washington DC. "We're probably good for the next 20 years for foreseeable missions," he says.

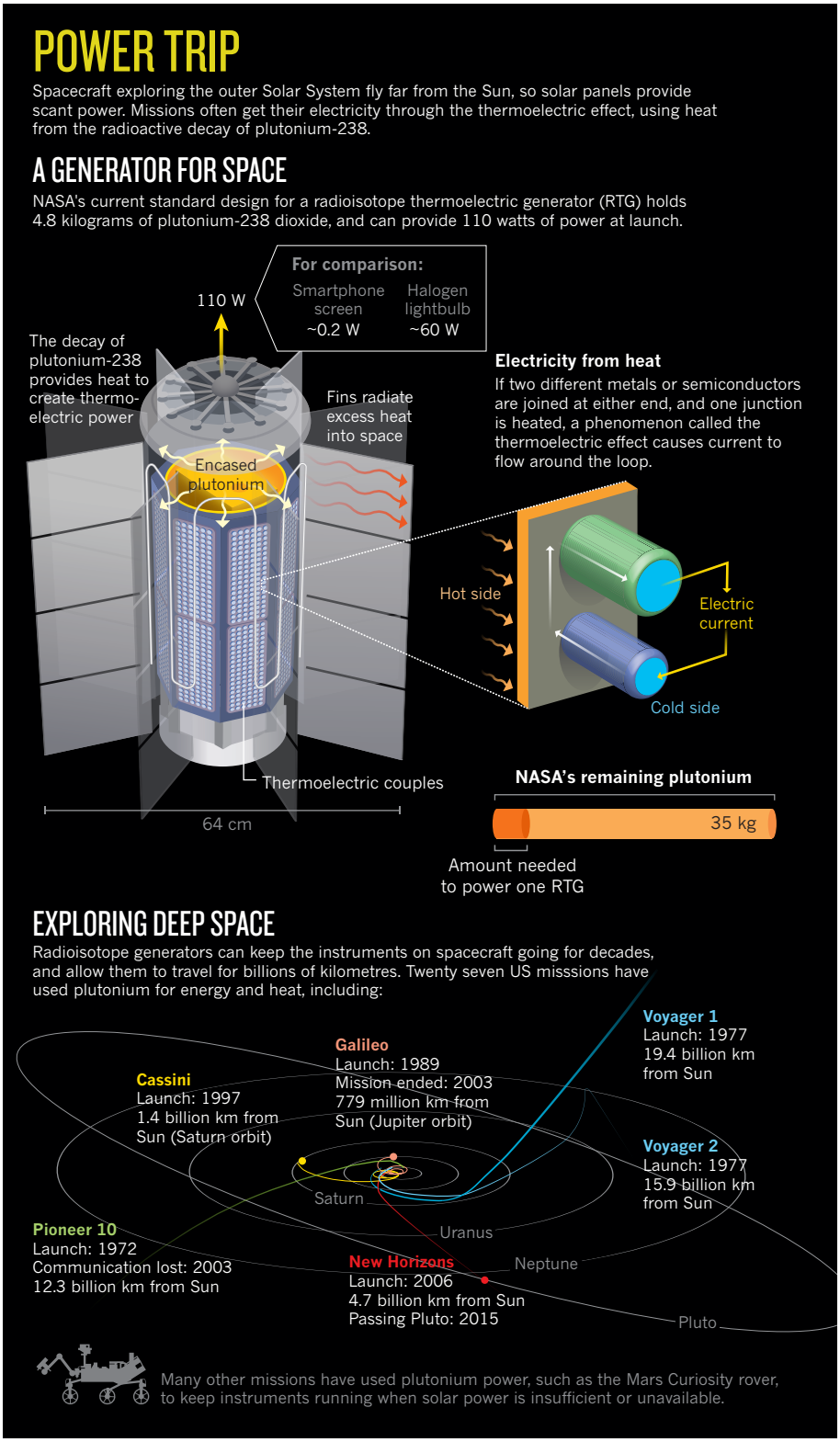
PRODUCTION LINE

The new ^{238}Pu production line starts at the Idaho National Laboratory in Idaho Falls, where the isotope neptunium-237 is chemically extracted from spent nuclear reactor fuel (see 'Fuel cycle'). The neptunium is then sent to Oak Ridge, the once-secret city where uranium was enriched for the first nuclear bombs during the Second World War. On a glorious Appalachian autumn morning, as reds and oranges begin to tinge the oak trees that give the region its name, it is easy to forget this nuclear history. But not for long: the road to the laboratory winds past the old uranium-enrichment plant and abandoned guard towers from the 1940s, which stand on either side.

On the ORNL campus, the ^{237}Np metal from Idaho is first pressed into pellets about the size and shape of a pencil eraser. The pellets are then slid one at a time into long aluminium tubes and taken to one of the lab's most historic buildings: the High Flux Isotope Reactor, the 49-year-old home of the highest neutron flux in the Western Hemisphere.

Irradiations manager Chris Bryan stands in an overlook area above what looks like an indoor swimming pool, showing off a miniaturized physical model of the reactor core assembly. It nestles in a cylinder of beryllium, 2.4 metres across and studded with dozens of holes. Before a typical reactor run, Bryan will slide each neptunium-filled tube into one of the holes, so that it is fully exposed to

NIK SPENCER/*NATURE*



the reactor core. "We're trying to squeeze as much neptunium into a finite volume as we can," Bryan explains. Many other nuclear- and materials-science experiments compete for the same space in the reactor.

Once the tubes are in place, Bryan will lower the whole assembly into the swimming pool, where the water will serve as a radiation shield, then switch the reactor on for 25 days. During that time, so many neutrons bombard the ^{237}Np that 10–12% of the nuclei in the sample absorb one. The result is neptunium-238, which quickly decays into ^{238}Pu.

Once this process is complete, the tubes are removed and taken next door, using a protected rail carriage, to the low-profile building where Wilson and his co-workers peer through their yellow windows and work their manipulator arms inside the lab's hot cells. Their job is to dissolve the irradiated pellets in nitric acid, then extract and concentrate the plutonium into an oxide powder that will eventually go into protective drums.

Finally, a radiation-shielded truck will drive the drums to the Los Alamos National Laboratory in New Mexico, where the oxide will be pressed into fuel pellets — although the laboratory will first have to replace its old, faltering pellet-pressing machine.

There are many other steps in this elaborate sequence. For one thing, Oak Ridge does not have enough space in its reactor to transform all of the ^{237}Np. Once the neptunium pellets are made there, some will be sent to the Idaho lab, whose Advanced Test Reactor will help out by doing some of the irradiation. Idaho will also store some of the finished plutonium pellets until they are needed for an MMRTG.

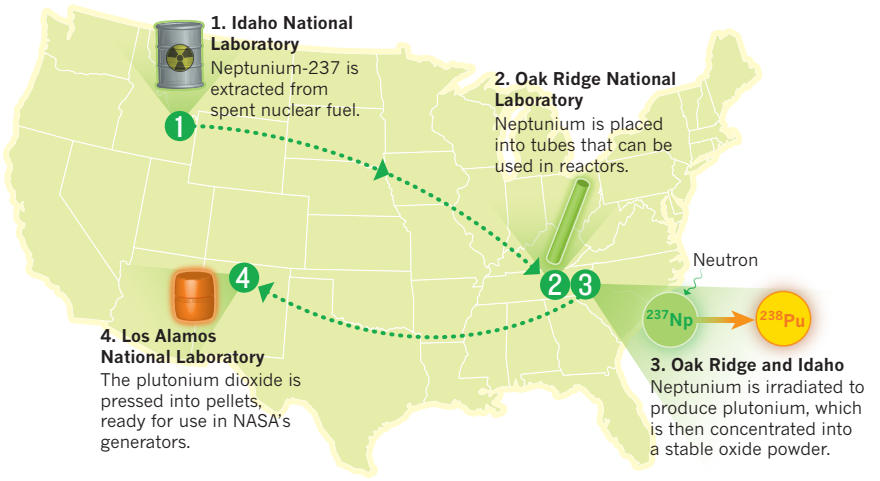
But for now the main focus remains in Oak Ridge. Robert Wham, a chemical engineer at the lab, is in charge of working out how to safely go from making a couple of test batches to churning out plutonium dioxide at the rate of 1.5 kilograms a year. Wham is the sort of quietly confident engineer whose eyes light up when he thinks about challenges such as designing an automated neptunium pellet-feeder, or picking the best length for the tubes to go in the Oak Ridge and Idaho reactors. "The people here hadn't worked with neptunium before," he says. "We're starting pretty much from scratch."

Now the major challenge is figuring out how to process all that plutonium in the limited number of hot cells at the Oak Ridge lab. Hot-cell technicians are in great demand; Wham will have to double the number of trained staff in the coming years. Already, the cells operate 24 hours a day, 7 days a week as the team works through test runs. "We're going great guns," Wham says. "Everyone wants to see this happen."

NASA is also looking at ways to extract more power from the plutonium it already has. At its Jet Propulsion Laboratory in Pasadena, California, materials engineer Jean-Pierre Fleurial leads a group that is exploring ways to build

FUEL CYCLE

Producing plutonium-238 for NASA's space missions will involve government laboratories across the United States.



thermocouples, the devices that generate electricity from plutonium's radioactive decay. By replacing the lead-based material currently used in the thermocouples with a cobalt–antimony material known as skutterudite, Fleurial's team will try to get at least 25% more power out of a generator at the beginning of its life. And this 'enhanced MMRTG' would also conserve power over time, which might substantially lengthen the lifetime of a spacecraft. It should be ready by 2022, Fleurial says.

POWER HUNGRY

Until last year, NASA was also working hard on space-going Stirling engines, which could generate as much power as an MMRTG from just one-quarter the amount of plutonium. Stirling converters work something like high-tech steam engines: the heat generated by plutonium decay drives the expansion of helium gas, which in turn moves a set of pistons to provide power. Missions enabled by Stirling technology might have included a boat to sail on the lakes of Saturn's moon Titan, or a 'comet hopper' that can manoeuvre to different places on a comet's surface. But NASA cancelled the programme in November 2013, citing cost constraints.

The decision sparked criticism from planetary scientists such as Jessica Sunshine at the University of Maryland in College Park, who is frustrated by what she sees as a lack of long-term planning for how to deal with NASA's limited plutonium supply. For example, NASA's latest call for mission proposals — for relatively low-cost Discovery-class spacecraft — does not even allow the use of radioisotopes for anything other than minimal heating of instruments. "How are we getting from DOE's restarting the programme to NASA's flying something?" she asks in reference to the plutonium supply. "Where is that path and how long is that going to take?"

Despite the agency's decision to cancel the Stirling programme, a small research effort has continued. John Hamley, manager of the radioisotope power systems programme at NASA's Glenn Research Center in Cleveland, Ohio, and his team have continued studies on 12 Stirling convertors in various configurations, which have been running for as long as 10 years. The aim is to prove that the pistons can work reliably for the long periods of time needed during an extended space mission.

All these efforts to conserve plutonium and produce more of it may still not be enough if NASA needs the isotope to power human exploration of space. The agency is now talking about sending astronauts to an asteroid or beyond, something that will require much more power than can be supplied by small chunks of ^{238}Pu. Whereas a planetary mission might require 300–900 watts of power, the much larger spacecraft needed for human deep-space exploration would require several tens of kilowatts, Schurr says. An internal NASA report, due out early next year, has been evaluating the needs for nuclear power in space. It may well conclude that it needs a self-sustaining power source, such as a fission reactor, which the United States has not used in space since 1965.

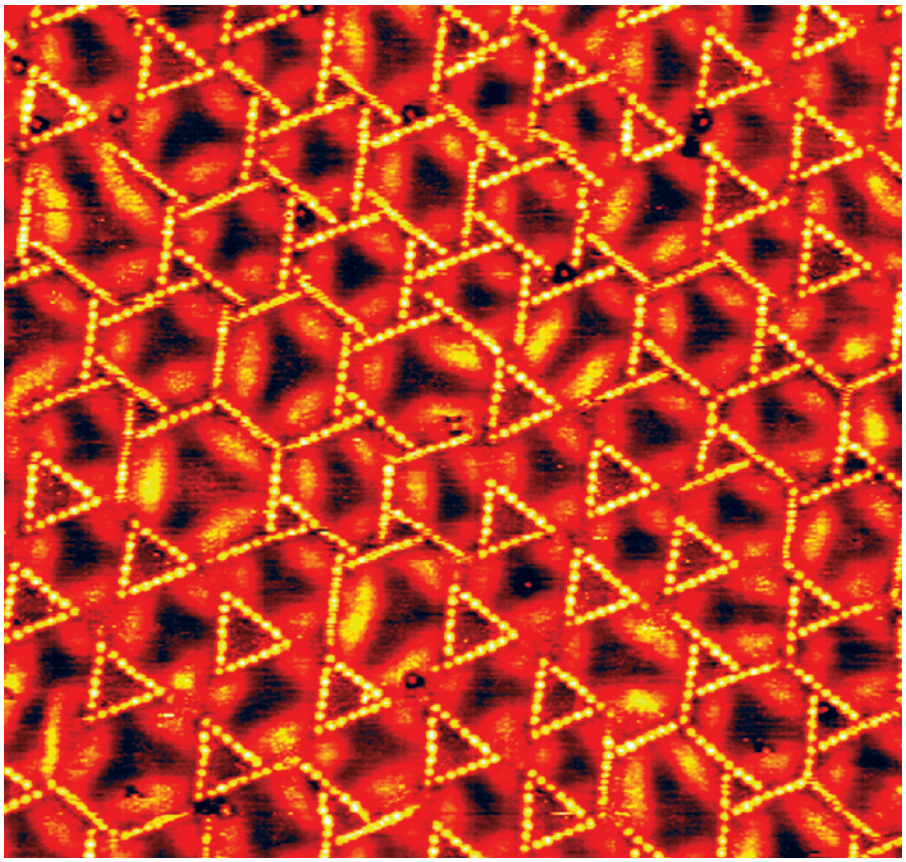
Back at Oak Ridge, Wham is thinking about how to make more plutonium, too. He leads the way through a narrow plywood-lined passageway in another building on the campus, which emerges into a cavernous concrete hall. The room was constructed for additional hot-cell space back in the 1960s, when the DOE was considering building nuclear reactors that run on thorium. Those plans have long since been shelved, but the well-shielded workspace remains.

If need be, Wham says that he could fashion more hot cells here and make even more plutonium — and find a use for the room in this chapter of atomic history, even if it did not find one in the last. "If they do come to us and want more," he says, "we know how to do it." ■

Alexandra Witze *writes for* Nature *from Boulder, Colorado.*

COMMENT

BROOKHAVEN NATL LAB./SPL



Sulphur atoms (yellow) 'dance' on a copper-layered catalyst under a scanning tunnelling microscope.

Hasten high resolution

Build precision microscopes to map atoms, say **Stephen J. Pennycook** and **Sergei V. Kalinin**.

The best electron and scanning probe microscopes today can resolve individual atoms and chemical bonds[1–4]. Views of materials such as graphene, catalysts and oxides on these scales — around 0.5 ångströms — reveal structures and the impacts of crystal defects on their properties.

To truly understand materials' chemical and physical properties, atomic arrangements need to be mapped with much greater precision. Resolutions of 0.1 Å — the goal set by physicist Richard Feynman in his 1959 American Physical Society lecture, 'There's Plenty of Room at the Bottom'[5] — would take us to the physical limit of microscope vision, set by the thermal vibrations of atoms. Small structural distortions that determine magnetism, valence (the number of chemical bonds an atom can form) and spin state would become apparent.

Currently, limits inherent to electron-microscope optics restrict us to seeing atoms or columns of atoms in two dimensions. Lens imperfections, electronic instabilities, thermal noise and environmental factors also blur views. Some scientists argue that microscopes will never clear these hurdles[6,7]. Others feel that because there are no materials in which atoms are spaced closer than 0.5 Å, greater resolution is not worth chasing.

We disagree. Keener microscopes are needed urgently to solve major world problems: solar, battery and fuel cells, computer memory chips and solid-state lighting all need to be more efficient. Three-dimensional (3D) maps of atoms would reveal how their interactions enable or limit functionality and, importantly, how materials can be improved.

Within a few years, at relatively low cost, researchers could improve microscope resolutions to 0.2 Å by honing aberration-corrector designs that are already available[8]. Stability could be improved through judicious tweaks to microscopes, materials, optics and electronics, and by reducing ambient noise. The main barrier is commercial — companies that build microscopes do not invest in specialist technologies if demand, and thus financial return, is expected to be low.

Three things are needed to accelerate microscope technology development: partnerships between academia and industry; government seed funds; and centres of excellence to develop computing power, data storage and analytical techniques.

CLEAR VISION

Exceeding atomic resolution is crucial for understanding important classes of materials such as superconductors, magnets and catalysts. It is often the small deviations from symmetry in atom positions that allow materials to store charge, information or energy — for example, in ferroelectric oxides used in computer memory chips or electrocatalytic oxides used in solid fuel cells. The complex atomic arrangements ▶

of nanophase metals, ceramics, alloys, solar cells, batteries and different types of glass have yet to be probed.

Interfaces between different materials — such as magnet–superconductor or oxide–oxide junctions — might exhibit properties such as electrical conductivity, chemical reactivity, superconductivity and ferromagnetism that are not found in their separate constituents. More-exact measurements of bond lengths and angles, ideally in three dimensions, are needed to hone materials for use in next-generation energy and information-technology devices.

Aberrations are inherent in electron lenses, which use magnetic fields and which, unlike glass lenses, cannot be shaped to arbitrary curvature. As in a camera, opening the aperture reduces the depth of field but also increases depth resolution. At today's practical resolution limit of 0.5 Å, the limited apertures available restrict depth resolution to the nanometre scale, which is too coarse to discern individual atoms.

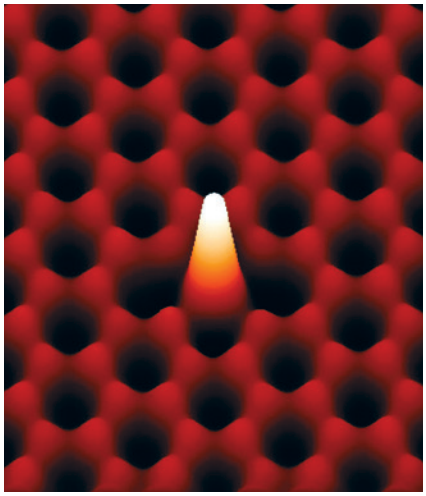
Lateral resolution at Feynman's level would allow us to distinguish atoms vertically. No longer would specimens need to be aligned to their internal crystal planes. One image would reveal a cross-section. A series of images taken using different foci would build up a 3D scan. Seeing atomic positions in three dimensions could distinguish between competing theories for a material's behaviour.

Such scans, like today's imagery, would have to be limited so that the sample is not destroyed by the electron beam. And biologists will have to take images using shorter, lower-current scans than can be used for materials. Researchers will need to learn new tricks, but the pay-off would be huge. As resolution improves, noise reduces, and we will be able to scan faster and monitor how things change with time.

"Exceeding atomic resolution is crucial for understanding important classes of materials."

Other types of microscopy would also be improved with higher resolution. Electron energy loss spectroscopy (EELS) would benefit from 3D capability to reveal elements, chemical valence and energy band levels at the same time as atomic structure.

In scanning probe microscopy (SPM), an image forms as a result of interaction between a sharp probe tip and the sample surface. Measuring the current between the two, as in scanning tunnelling microscopy, or the minute forces, as in atomic force microscopy, traces the structure of the surface. Although the maximum resolution achievable is limited by the fundamental physics of tip–surface interactions, new low-noise systems will allow mapping of displacements of surface atoms and changes in bond lengths to less than 10 picometres.



WU ZHOU/OAK RIDGE NATL LAB.

A scanning transmission image reveals a silicon atom in a layer of graphene.

By probing electronic, phonon (vibrations in a lattice) and spin responses with SPM, we will better understand the factors that control the ferroelectric, magnetic and superconductive functions of materials. By tuning the force and current applied through the probe tip, atoms and molecules could be manipulated and their chemical and electrochemical responses explored away from equilibrium.

Improved energy resolution would enable physicists to map energy band gaps and phonons in materials for solid-state lighting, thermoelectrics and solar cells using EELS[9]. This resolution could be achieved using advanced electron optics such as monochromators that narrow the energy spread of the electron beam from today's 300 millielectronvolts to 10 millielectronvolts or less. SPM provides information on local superconductivity, energy band gap and molecular vibrations or phonon structure.

For electron and scanning-probe microscopy, additional signals, such as emitted light or electronic current, could be collected simultaneously. From this, one could test whether particular lattice defects kill or enhance the effectiveness of solid-state lighting or solar cells, how a molecule interacts with the substrate, or how local polarization gradients affect oxidation states and magnetic properties in ferroelectrics and polar materials.

PUSHING AHEAD

To achieve Feynman's goal, microscope optics and electronic and mechanical stability must be improved. We need new designs for correctors with larger apertures.

A main problem is a lack of awareness among scientists of how much can be gained from even higher-resolution microscopy. With aberration correctors selling well (hundreds each year), there is little incentive for manufacturers to develop new capabilities. The low-noise scanning probe microscopes used for cutting-edge studies are largely lab-built.

Significant further investment will be necessary to deliver a factor-of-two improvement in spatial resolution in electron microscopes within five years, just as multimillion-dollar government-funded projects in the United States and Japan have led to the previous factor-of-two resolution increase in the past decade. As was the case with today's aberration-corrected machines, the new state-of-the-art microscopes would soon become available, at a probable cost of between US$5 million and $10 million each.

One of the pathways to achieving this goal is through community-wide workshops to construct a road map for instrumental developments and identify scientific opportunities. The crucial transition from taking images to acquiring detailed information on atomic positions, bond lengths and local functionalities will require new methodologies. Large, multi-dimensional data sets will pose challenges for data collection, storage and analysis. New approaches will be needed for extracting relevant knowledge and linking it to theory.

The scientific community should set up centres to host and coordinate the high-power computing services needed to support high-resolution microscopy. Shared online environments will foster collective interpretation. By pooling data, fewer experiments would have to be repeated, providing the experimental counterpart to other programmes sharing analytical tools, such as the $100-million US Materials Genome Initiative.

To paraphrase Feynman: there's still plenty to see at the bottom. ■

Stephen J. Pennycook *is research professor in the department of materials science and engineering, University of Tennessee, Knoxville, Tennessee, USA.* **Sergei V. Kalinin** *is a director of the Institute for Functional Imaging of Materials, and theme leader at the Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA.*
e-mail: spennyco@utk.edu

1. Erni, R., Rossell, M. D., Kisielowski, C. & Dahmen, U. *Phys. Rev. Lett.* **102,** 96101 (2009).
2. Sawada, H. *et al. J. Electron. Microsc.* **58,** 357–361 (2009).
3. Krivanek, O. L. *et al. Nature* **464,** 571–574 (2010).
4. Zhou, W. *et al. Phys. Rev. Lett.* **109,** 206803 (2012).
5. Feynman, R. P. *J. Microelectromech. Sys.* **1,** 60–66 (1992).
6. Reich, E. S. *Nature* **499,** 135–136 (2013).
7. Uhlemann, S., Müller, H., Hartel, P., Zach, J. & Haider, M. *Phys. Rev. Lett.* **111,** 046101 (2013).
8. Sasaki, T. *et al. J. Electron. Microsc.* **59,** S7–S13 (2010).
9. Krivanek, O. L. *et al. Nature* **514,** 209–212 (2014).

A decommissioned missile in a silo at the Titan Missile Museum in Green Valley, Arizona.

JAMES MARSHALL/CORBIS

MILITARY SCIENCE

Scientific spoils of war

Ann Finkbeiner examines two books on the cold war's ethical and material legacies.

War is good for science. Countries require their defence industries to invent military technologies, which are often based on science, sending money to researchers. So how does this intersection affect the course of research? Two books discuss the extent to which scientists change — or must change — what they do in response to national emergencies.

The cold war is an excellent case study. It saw the continuation of the extraordinary development of nuclear weapons, ballistic missiles and radar begun during the Second World War. *Science and Technology in the Global Cold War*, an essay collection edited by science historians Naomi Oreskes and John Krige, addresses the question: were scientists guided by curiosity, or did national funding redirect them towards military technological applications? Its answer: although redirection is inevitable and powerful, so is curiosity.

The balance differed from field to field and place to place. China and the Soviet Union erased the distinction between pure and applied science and directed their researchers towards national priorities — an isolated self-reliance in China, and big industry in the Soviet Union. In China, horse breeding was given the status of scientific experimentation; in the Soviet Union the Sputnik satellite, launched in 1957, was deemed "Soviet

Science and Technology in the Global Cold War
EDITED BY NAOMI ORESKES AND JOHN KRIGE
The MIT Press: 2014.

Unmaking the Bomb: A Fissile Material Approach to Nuclear Disarmament and Nonproliferation
HAROLD A. FEIVESON, ALEXANDER GLASER, ZIA MIAN AND FRANK N. VON HIPPEL
The MIT Press: 2014.

science". But the balance never stayed put. Chinese researchers and students looked towards international science: between around 1980 and 2000, at least 10,000 went abroad to work and study. Soviet nuclear scientists outmanoeuvred the state and, by relabelling pure science as applied, succeeded in creating a reactor design based more on technical feasibility than on cheapness.

The West experienced a similar shifting balance. The US response to Sputnik led to NASA's Apollo human-spaceflight programme, but as that project slowly ground down, NASA adopted its technologies for the Mission to Planet Earth observation system, which gathered data for climate scientists. Radar, developed by Britain and the United States to track aircraft and missiles, was used in the 1960s by US physicist Irwin Shapiro to test Einstein's general theory of relativity. A cold war US surveillance system that used underwater sound recordings to trace the movements of submarines was recycled around the year 2000 by scientists at the Scripps Institution of Oceanography in San Diego, California, to map ocean temperatures and global warming. If national priorities bend science towards application, scientists bend it back towards pure research.

Written mostly by historians of science, *Science and Technology in the Global Cold War* is an academic conversation with no grand conclusions. But one commonality that emerges, writes Krige, is that "he who paid the piper didn't so much call the tune as provide the instruments, the hardware, and the logistical support". Changing the metaphor, national attempts to direct science look like a magnetic field aligning iron filings — until the filings go off on their own in all directions.

Unmaking the Bomb presents a more complex relationship between scientists and war, arguing that researchers tasked with creating extraordinarily lethal applications have a responsibility to control them. Specifically, the authors — physicists and nuclear-policy experts Harold Feiveson, Alexander Glaser, Zia Mian and Frank von Hippel — present the case for controlling the materials that make a nuclear bomb nuclear.

National mandates drove the nuclear bomb's development during the 1940s. By ▶

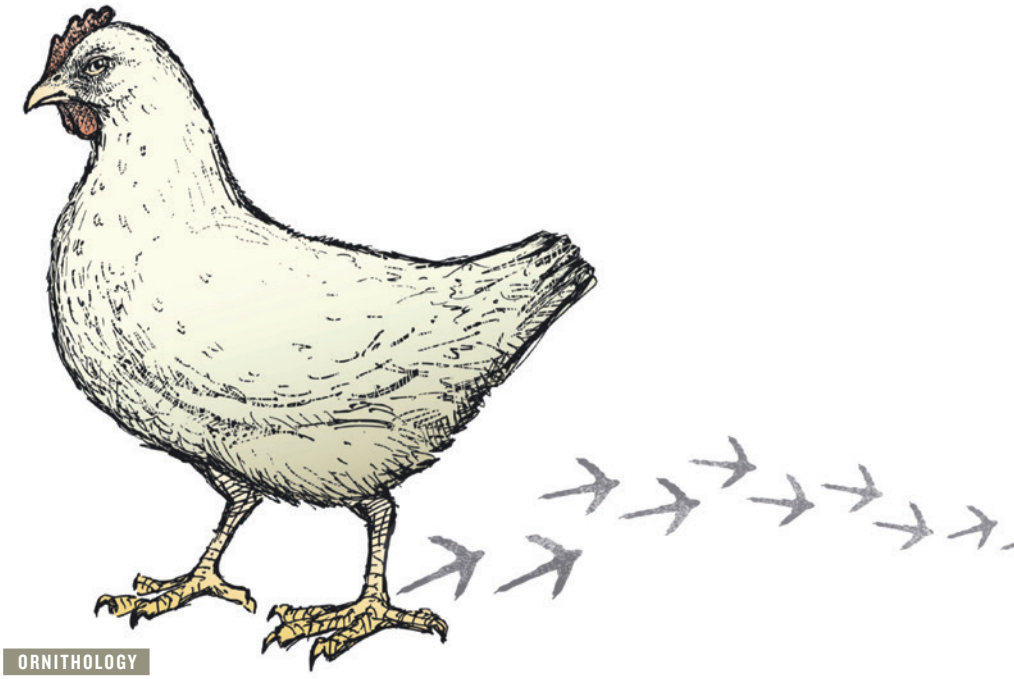
the peak of the cold war, 10 countries — including the United States, the Soviet Union, Britain and China — had built 65,000 nuclear warheads. But before the first bomb had been built, nuclear scientists had been lobbying politicians to change the mandate from building nuclear weapons to controlling them. The lobbying, partly through avenues such as the Pugwash Conferences on Science and World Affairs, was fairly successful: the number of nuclear weapons in those 10 countries has fallen to around 17,000.

But the fuel — fissionable plutonium or uranium enriched in a rare isotope of uranium — is still with us. Neither occurs naturally, so bomb-builders manufactured them. At the end of the Second World War, 100 kilograms of weapons-grade material had been made; now, it is 1,900 tonnes, enough for 100,000 bombs. As the authors show, material from dismantled bombs can be downblended to a less fissionable form and stored or used in power plants, but it cannot be destroyed, and it remains available for nuclear weapons or for low-tech radiological weapons. In 1945, only the United States could build a nuclear warhead; now, 35–40 countries can, and the margin of security is "too slim for comfort", says a former director-general of the International Atomic Energy Agency.

Feiveson, Glaser, Mian and von Hippel convincingly argue that this problem demands a real and immediate solution. Along with the history of nuclear weapons, they cover attempts to control the weapons' spread, including the 1970 Treaty on the Non-Proliferation of Nuclear Weapons; the physics and technology of producing, downblending and storing fuel; and the complexities of convincing nations to agree to be supervised and controlled by an international agency.

The authors' suggested long-term policy is to reduce the amount of fissionable material in military and civilian stockpiles, and to regulate it "as if the world is preparing for complete nuclear disarmament". Countries should stop hiding the sizes of their stockpiles, the authors write, and stop manufacturing weapons-grade uranium and plutonium; they should also downgrade or bury all fissionable material, even if they must give up nuclear energy. Finally, they should agree to international verification of declarations about weapons production — even if that means relying on nuclear scientists rather than politicians to tell the truth. ■

Ann Finkbeiner *is a freelance science writer in Baltimore, Maryland, who often covers scientists advising governments. She blogs at* The Last Word on Nothing, *www.lastwordonnothing.com.*



DZHAMILIYA ERMAKOVA/GETTY

ORNITHOLOGY

Fowl domination

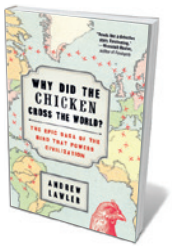
Ewen Callaway relishes a study tracing the chicken's eventful march from Asian jungles to global ubiquity.

The chicken is the Swiss army knife of livestock. Since its domestication in Southeast Asia as early as 18,000 years ago, the bird has been religious sacrifice, pet, research subject, fighting machine and, of course, dinner. The Victorians paid enormous sums for exotic breeds, and in the 1960s, NASA imagined the birds feeding Martian colonies. Around 20 billion are alive at any one time, bred to meet global demand. Science journalist Adrian Lawler explores the chicken's multipronged place in human civilization in his rip-roaring, erudite *Why did the Chicken Cross the World?*

Genome data and resemblance have pinpointed the red jungle fowl *Gallus gallus* — a furtive bird that roams the subtropical forests of southern Asia — as the wild ancestor of *Gallus gallus domesticus*. The birds are considered one species, because unions between them still produce fertile offspring. A few thousand years of separation is an evolutionary blink of the eye, too brief to create reproductive barriers.

Scientific efforts to unpick the origins of the domestic chicken are muddied by the fact that few, if any, living red jungle fowl are free of the genetic vestiges of their ancestors' romps with domestic chickens. The last purebred jungle fowl on Earth may reside, as Lawler shows, on a farm in the northeast of the US state of Georgia, rather than in a forest in Malaysia.

That is down to ornithologist Gardiner Bump. In the 1950s and 1960s, faced with a shortage of game birds in the US southeast, Bump set out to populate forests with imported wild red jungle fowl. He paid trappers to collect eggs — the more remote the better, because he wanted purebred birds — and deliver them to US hatcheries. The birds never thrived, and the US government pulled the plug on the programme in 1970. Descendants of Bump's birds survive in a handful of flocks. An evolutionary geneticist has sampled their blood, in the hope of discovering what truly sets chickens apart from their wild forebears.



Why Did the Chicken Cross the World?: The Epic Saga of the Bird That Powers Civilization
ANDREW LAWLER
Atria: 2014.

From their initial domestication, Lawler traces the chickens' journey to Mesopotamia and ancient Egypt, where the earliest known depiction of the bird was made, and then on to Polynesia and South America, where DNA from ancient chicken bones offers contentious evidence for a pre-Columbian trans-Pacific chicken trade. The author does not dwell on such controversy for long. For much of the book, science has a supporting role to history, ethnography and even advocacy.

Lawler's discussion of cockfighting is among the book's most compelling material. In ancient Greece, Babylon and China, pitting roosters against each other was embedded in religious practice. Now mostly illegal, it still thrives in parts of South America and Asia, especially the Philippines, as Lawler demonstrates with a harrowing dispatch

from the World Slasher Cup in Manila. He shows cockfighting as the brutal pastime it is, while recognizing it as an important chapter in human–chicken relations.

Chicken's mealtime ubiquity dates from the twentieth century. African Americans and Jewish immigrants brought the bird into US cities, and farmers who had once viewed chicken-keeping as women's work survived the Great Depression thanks to income from the birds. But wartime rationing of other meat put chicken on every plate. First held in 1948, the US Chicken of Tomorrow contest was conceived by supermarket chain A&P (and later sponsored by the US Department of Agriculture) to improve the efficiency of poultry production and expand the fledgling market. Before the contest, chickens bred for meat took 70 days to reach an average of 1.4 kilograms. Modern birds take 47 days to reach 2.6 kilograms, and they convert feed to meat 50% more efficiently (although many spend their lives in chronic pain because of the extra body mass). US chicken consumption is now four times what it was before the contest.

Readers of Michael Pollan's *The Omnivore's Dilemma* (Penguin, 2006) or Christopher Leonard's *The Meat Racket* (Simon & Schuster, 2014) will know the rest of the story. Leonard used the term "chickenization" to describe the 'vertical integration' of meat production developed and perfected by conglomerates such as Tyson Foods, whereby farmers have no ownership or control over the flocks they breed, which often number tens of thousands of birds. Americans eat more chicken meat per capita than any other nation, but the rest of the world is catching up. China surpassed the United States in overall chicken consumption in 2012. Meanwhile, the mass culling of chickens across Asia to stop an avian-influenza pandemic shows that chicken health is a global concern.

Lawler is not the first to denounce the inhumane treatment of the animals or to raise the red flag about bird flu. But his perspective as a science reporter gives fresh insight into the problems created by the ubiquity of chickens — as well as possible solutions. Especially compelling is the profile of Janice Siegford at Michigan State University in East Lansing, who is studying how to improve the welfare of chickens bred for food ('cage free' labelling is no guarantee that a chicken does not suffer throughout its life). Lawler recognizes that modern chickens — perhaps unlike genuine red jungle fowl — are here to stay. Who knows, maybe they will one day make it to Mars. ■

Ewen Callaway *writes for* Nature *from London.*

CLIMATE SCIENCE

A climate trance

Richard Van Noorden considers a technical lecture that ultimately fails as theatre.

House lights down. A spotlight picks out a man, seated: climate scientist Chris Rapley. "I'm here to talk about the future," he says. Behind him on three giant video walls swirl greyscale images of tides and seas, and satellite views of Earth. So begins *2071*, a piece about climate change at London's Royal Court Theatre.

Rapley calmly lays out his credentials. Professor of climate science at University College London; former director of the British Antarctic Survey; former director of London's Science Museum. At a measured pace, he unfolds what he has seen and what scientists have learned, through means such as satellites, ocean buoys and ice cores, about the crumbling West Antarctic Ice Sheet, sea-level rise, the Holocene and Anthropocene epochs, and the interactions between lithosphere, biosphere, hydrosphere, cryosphere and atmosphere. The grey backing visuals break into big, moving white-on-black bar charts.

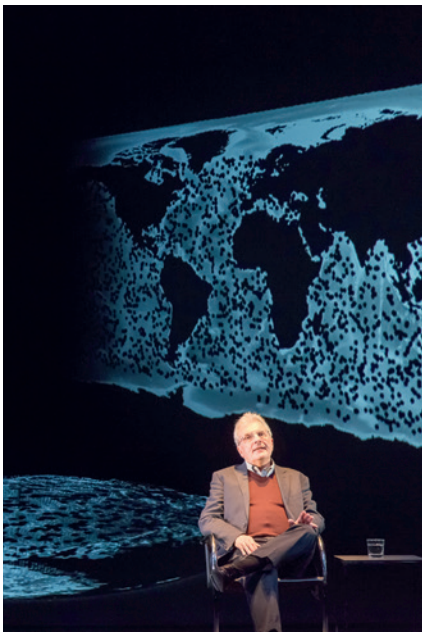
After 15 minutes, the audience realizes that there will be no let-up: *2071* is not a play, but an address just over an hour long. Rapley is the sole performer. This is a scientific lecture.

Global sea-levels are rising by 3.3 millimetres a year, Rapley says. The ocean is acidifying. Changes in solar radiation are not responsible for the observed temperature rise, because we see that the upper atmosphere is not warming, but cooling. The multisyllabic drone goes on, a flow of data lent emotional resonance only by a tense, unsettling soundtrack.

2071
WRITTEN BY DUNCAN MACMILLAN AND CHRIS RAPLEY; DIRECTED BY KATIE MITCHELL
Royal Court Theatre, London. 5–15 November 2014. The Deutsches Schauspielhaus, Hamburg, Germany. 17–18 December 2014.

Rapley and director Katie Mitchell are trying, perhaps in response to the histrionic climate politics of recent years, to establish a quiet, concentrated atmosphere in which to lay out the facts. But Rapley's monochrome recital risks sending his viewers into a climate trance, eyes glazed over by science. At one point, he starts quoting verbatim from the latest report of the Intergovernmental Panel on Climate Change; his delivery hardly changes in tone. The script's worst sin is that it fails even on its own terms. Although he sets himself up as bringing home scientific truths, Rapley in fact makes no effort to convey the human realities of acidifying oceans, rising sea levels, or two- or four-degree rises in global temperature. (The play's title makes a stab at humanizing the proceedings — *2071* is the year when Rapley's eldest grandchild will be the age he is now. But it's an awkward attempt.)

Mitchell's critically acclaimed 2012 play on population, *Ten Billion*, featured another professor, computational scientist Stephen Emmott, delivering another talk, in a stage recreation of his office. But where that was an entertaining polemic ("I think we're fucked," Emmott concluded), *2071* is sober and technical. Despite the scientific consensus behind Rapley's words, it is difficult to imagine that it will engage even a willing audience. Aiming for authenticity, Mitchell and Rapley have missed a chance to create a piece of drama that really gets under the skin of the issue; one that might seamlessly blend instruction and inspiration. But for those with an appetite for the stark facts on climate change, *2071* is just the ticket. ■



STEPHEN CUMMISKEY

Climate scientist Chris Rapley in *2071*.

Richard Van Noorden *is a senior reporter at* Nature.

Correspondence

Ebola: models do more than forecast

Your assertion that models of the Ebola epidemic have failed to project its course misrepresents their aims (see *Nature* **515,** 18; 2014). They helped to inspire and inform the strong international response that may at last be slowing the epidemic (see M. F. C. Gomes *et al. PLoS Curr. Outbreaks* http://doi.org/vvd; 2014).

Subsequent models assessed the likely impact of different public-health interventions and policy decisions (J. A. Lewnard *et al. Lancet Infect. Dis.* **14,** 1189–1195 (2014) and A. Pandey *et al. Science* http://doi.org/wts; 2014). As those interventions were implemented and as people's behaviour changed, case counts below the modelled baseline were early indicators that the response to the outbreak was having an effect.

Epidemics are affected by countless variables, so uncertainty is a given. Models synthesize available information. Without them, there is little to guide decision-makers during an outbreak. Their importance goes beyond providing forecasts.

Caitlin Rivers* *Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, USA.*
cmrivers@vbi.vt.edu
**On behalf of 24 correspondents (see go.nature.com/dfzbde for a full list).*

Ebola: the power of behaviour change

Without including social, cultural and behavioural responses to the Ebola epidemic, models may overestimate outbreak size (*Nature* **515,** 18; 2014).

Behavioural response, triggered by an epidemic, can slow down or even stop virus transmission (see S. Funk *et al. Proc. Natl Acad. Sci. USA* **106,** 6872–6877; 2009). Indeed, altered cultural perception in response to the disease enabled people's behaviour to change in ways that helped to contain outbreaks in the past (see B. S. Hewlett and R. P. Amola *Emerg. Infect. Dis.* **9,** 1242–1248; 2003).

Reports from Foya in Liberia indicate that the outbreak there is now in decline. A local information campaign to change funeral practices and other behaviours seems to have paid off.

More aid and more personnel are urgently needed, but so is the involvement of local communities and the provision of information that can help to contain this epidemic.

Sebastian Funk, Gwenan M. Knight *London School of Hygiene & Tropical Medicine, London, UK.*
Vincent A. A. Jansen *Royal Holloway University of London, Egham, Surrey, UK.*
vincent.jansen@rhul.ac.uk

Can brain training boost cognition?

Eminent scholars from around the world last month signed a statement on the 'brain training' industry (see go.nature.com/d2bpuj). They point out discrepancies between current scientific understanding of cognitive enhancement and advertising claims for commercial cognitive-training software. But it should not be inferred that software can never improve cognition (see D. Bavelier *et al. Nature Rev. Neurosci.* **12,** 763-768; 2011).

The effectiveness of 'brain-training' software may differ, so product claims might not always be exaggerated. And, given that new training programs are compared with results from control groups engaging in other stimulating activities, the absence of any effect after training can be relative, and is not necessarily definitive. Cognitive enhancement is a vast enterprise that has not yet been clearly defined, and finding optimal ways to train the brain is still a promising area of research.

David Moreau *Princeton University, New Jersey, USA.*
dmoreau@princeton.edu

Chinese universities: beware cronyism

Jie Zhang rightly points out that China's universities need high-quality faculty members if they are to be competitive internationally (*Nature* **514,** 295–296, 2014). But there are risks in giving individual colleges and departments more autonomy in recruiting staff.

Excessive recruitment has led to bloating and inefficiency in some Chinese universities, which might be exacerbated by greater autonomy. More recruiting freedom could also encourage cronyism, which already undermines research and higher education in China (see, for example, go.nature.com/vl4rxt; in Chinese).

To avoid these pitfalls and to ensure the quality and diversity of new faculty members, university panels should appoint the strongest academic performers from a shortlist drawn up by departmental recruiting panels in an open and transparent process.

Hong-Wei Xiao *China Agricultural University, Beijing, China.*
xhwcaugxy@163.com

Chinese universities: gear up for Nobels

We agree with Jie Zhang that university reform is needed to improve the quality of Chinese research papers (*Nature* **514,** 295–296; 2014). A home-grown scientist in China might then stand a chance of winning a Nobel prize for the first time.

Novelty in research remains rare: publications with a Chinese scientist as first or corresponding author accounted for just 362 papers in *Nature* and 388 in *Science* from 1992 to 2012. This is despite the government's increased investment in research during 2002–12 and its launching of talent schemes and large scientific programmes in past decades. China needs to recognize that Nobel prizes are awarded for scientific breakthroughs, not for short-term successes.

Switching the country's emphasis from publications in journals in Thomson Reuters' Science Citation Index to multiple evaluation criteria could help, and would offset academic corruption in promotions and student graduation.

China should implement fixed annual salaries for scientists, rather than paying their incomes as a component of research funding, which undermines motivation. The scientific administration system needs overhauling, particularly with respect to funding applications and individual performance evaluation.

Yi-Ping Chen, Yi-Shan Lin, Yi Zhang *Institute of Earth Environment, Chinese Academy of Sciences, Xi'an, China.*
chenyp@ieecas.cn

Rename comet probe after Greek hero

The Rosetta spacecraft's Philae probe, which landed successfully on an orbiting comet on 12 November (see *Nature* http://doi.org/w8k; 2014), could be renamed Pheidippides — for its record-setting marathon run and transmission of its message before collapsing.

Len Fisher *University of Bristol, UK.*
len.fisher@bristol.ac.uk

CORRECTION

The Outlook article 'The search for the rice of the future' (*Nature* **514,** S60–S61; 2014) wrongly stated that a flood-resistant gene was bred into rice by Pamela Ronald. In fact, the breeding was done by David Mackill, Abdelbagi Ismail and their team at IRRI.

CLIMATE SCIENCE

El Niño's variable history

A study of the El Niño phenomenon over the past 21,000 years suggests that El Niño responded in complex ways to a changing climate, with several competing factors playing a part in its varying strength. SEE LETTER P.550

JOSEPHINE R. BROWN

The episodic warming and cooling of the tropical Pacific Ocean's surface waters, known as the El Niño–Southern Oscillation (ENSO), is responsible for large year-to-year variations in global climate, and causes widespread effects that include droughts (Fig. 1), floods and fires. It is therefore important to know how ENSO may change in the future. To address this question, researchers can look to past climates for clues. In this issue, Liu *et al.*[1] (page 550) present a set of climate-model experiments investigating the evolution of ENSO over the 21,000 years since the peak of the most recent glacial period. Their experiments show that ENSO varied on a range of timescales, producing a complex history with epochs of weaker and stronger activity.

ENSO is a natural fluctuation of climate that arises from interactions between the atmosphere and ocean in the tropical Pacific, with El Niño events (warming) occurring about every three to seven years. The strength of these events may vary as a result of natural, internal feedback processes[2] as well as through external factors such as human-induced global warming[3]. Over the past 21,000 years, the global and regional climate changed in response to variations in incoming solar radiation driven by slow changes in Earth's orbit around the Sun; the melting of continental ice sheets; and natural fluctuations in greenhouse gases. In their study, Liu and colleagues consider how ENSO was modified by these competing factors.

Previous studies of past ENSO behaviour using complex global models have been restricted to simulations of 'time-slices' of key periods such as the Last Glacial Maximum (21,000 years ago) and mid-Holocene (6,000 years ago)[4]. These time-slices typically consist of only hundreds to a few thousand years because of the computer resources needed to run such experiments. Now, for the first time, Liu *et al.* use a complex global climate model — the Community Climate System model version 3 (CCSM3) maintained at the US National Center for Atmospheric Research — to simulate the full 21,000 years from the last glacial. This has allowed the sensitivity of ENSO to multiple influences (forcing) to be investigated in a continuously evolving climate. The model's forcing includes changes in Earth's orbital parameters, the concentrations of greenhouse gases and continental ice sheets, as well as ocean freshwater input from melting ice. The authors also carried out experiments to explore the sensitivity of ENSO to each of these factors individually.

ULET IFANSASTI/GETTY

Figure 1 | Severe drought. The El Niño–Southern Oscillation (ENSO) can cause severe drought in many parts of the world. Shown here is a villager walking through a field affected by one such drought in August 2009 in Lamongan, East Java, Indonesia. Liu *et al.*[1] have used climate models to explore the history of ENSO over 21,000 years, to examine how this phenomenon varies in response to a changing climate.

Liu *et al.* find that ENSO gradually became around 15% stronger during the Holocene (the past 11,000 years). This strengthening occurred in response to altered incoming solar radiation due to orbital changes, which led to warming of the tropics and increased feedbacks between the atmosphere and upper layers of the ocean. The role of orbital forcing is confirmed by the experiment that involved orbital changes only, which reproduces the gradual trend in ENSO amplitude.

The model simulates a slightly weakened ENSO at the Last Glacial Maximum. In the subsequent 'deglacial' period, as the climate warmed and continental ice sheets melted, ENSO amplitude varied on millennial timescales as a result of changes in ocean freshwater input from melting ice, which modified the circulation in the Atlantic Ocean. Liu *et al.* identify a set of mechanisms for this response, ultimately pointing to the varying magnitude of the annual cycle of temperature in the equatorial Pacific, which is strong when ENSO is weak and vice versa[5]. During the same period, increasing atmospheric carbon dioxide concentrations tended to weaken ENSO, whereas the impact of retreating continental ice sheets on atmospheric circulation led to a strengthening.

The authors compare these results with reconstructions of ENSO variability from proxy records such as corals, mollusc shells and lake and ocean sediments[6–10]. Several proxy records from the early to mid-Holocene (11,000 to 5,000 years ago) indicate that ENSO was weaker than it is now, with reductions of around 30–50%, although the timing of the minimum varies between records[6–9]. Another reconstruction does not show a significant change in ENSO variability at this time[10]. Climate-model simulations of the mid-Holocene[4] show a 10–15% weakening of ENSO (relative to the pre-industrial climate), which is consistent with Liu and co-workers' results but smaller than most proxy records suggest. The apparent underestimate of the change by models may reflect limitations of the proxies, or may highlight insufficient model sensitivity.

There are fewer proxy records of ENSO from the period before the Holocene, and the available records do not provide a clear picture of ENSO in glacial and deglacial climates[6,9]. Model simulations of the Last Glacial Maximum also fail to show a consistent change in ENSO variability[4]. Liu *et al.* propose that this lack of agreement between models may be due to the competing effects of large continental ice sheets and low atmospheric CO_2 levels, which influence ENSO in opposing ways during this time. Disagreement between proxy records suggests that the spatial pattern of ENSO impacts across

the Pacific may also have changed.

A great strength of this study is its use of multiple simulations using individual forcing to confirm the role of each factor in ENSO changes. However, the results are based on a single climate model, and it is well known that models differ in their simulation of ENSO. In particular, models disagree about changes in ENSO strength in both past climates such as the mid-Holocene and the Last Glacial Maximum[4] and in projections of future climate with increased concentrations of greenhouse gases[3]. For this reason, it is imperative that the authors' 21,000-year experiments are repeated using other climate models, ideally including the full set of sensitivity experiments.

Another challenge is to develop proxy reconstructions of ENSO that can provide a clearer picture of past variability than is currently available, and to improve methods of comparing models and proxy records[10,11]. In particular, the large unforced natural variability of ENSO implies that records spanning many decades or longer may be required to identify changes in strength that are outside this natural range[2,10]. As Liu *et al.* conclude, to provide a robust basis for comparison between models and proxy reconstructions, an expanded set of proxy records is needed, particularly from the equatorial Pacific region, which is most sensitive to ENSO.

Liu and colleagues' study constitutes a major step towards understanding the complex history of this crucial phenomenon. At the same time, their results suggest that ENSO may respond in opposing ways to different regional and global influences, highlighting the challenge of predicting its future activity. ■

Josephine R. Brown *is at the Centre for Australian Weather and Climate Research, Bureau of Meteorology, Melbourne, Victoria 3001, Australia.*
e-mail: j.brown@bom.gov.au

1. Liu, Z. *et al. Nature* **515,** 550–553 (2014).
2. Wittenberg, A. T. *Geophys. Res. Lett.* **36,** L12702 (2009).
3. Collins, M. *et al. Nature Geosci.* **3,** 391–397 (2010).
4. Masson-Delmotte, V. *et al.* in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Stocker, T. F. *et al.*) Ch. 5, 383–464 (Cambridge Univ. Press, 2013).
5. Timmermann, A. *et al. J. Clim.* **20,** 4899–4919 (2007).
6. Tudhope, A. *et al. Science* **291,** 1511–1517 (2001).
7. Moy, C. M., Seltzer, G. O., Rodbell, D. T. & Anderson, D. M. *Nature* **420,** 162–166 (2002).
8. Carré, M. *et al. Science* **345,** 1045–1048 (2014).
9. Sadekov, A. Y. *et al. Nature Commun.* **4,** 2692; http://dx.doi.org/10.1038/ncomms3692 (2013).
10. Cobb, K. M. *et al. Science* **339,** 67–70 (2013).
11. Brown, J., Tudhope, A. W., Collins, M. & McGregor, H. V. *Paleoceanography* **23,** PA3202 (2008).

MAMMALIAN EVOLUTION

A beast of the southern wild

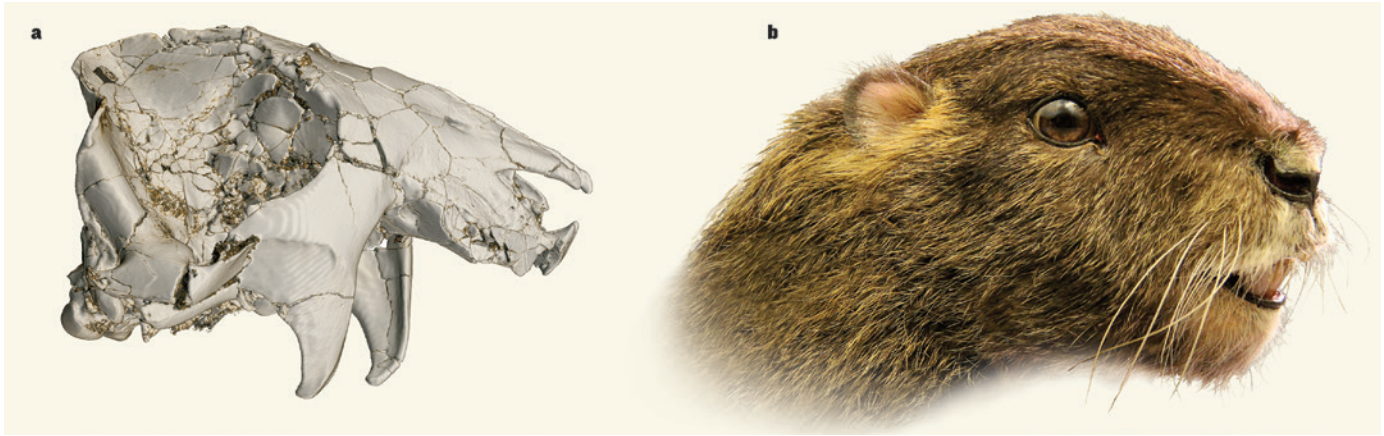
A newly discovered skull from the Cretaceous period belongs to a mammal that was big, strange and fast-moving. The fossil solves a long-standing mystery, and helps to resolve a controversy about mammalian evolution. SEE ARTICLE P.512

ANNE WEIL

What were the gondwanatheres? Even the palaeontologists who study them have been wondering for decades. Restricted to the southern continents, the Gondwanatheria were a mammalian oddity that have been found in mid-Cretaceous to early Eocene sediments from around 110 million years ago up to 45 million years ago. Until now, they were known only from the most fragmentary pieces — teeth here and there, and, rarely, a piece of mandible. Their distinctive, specialized, high-crowned cheek teeth revealed them to be omnivores and herbivores, but obscured their relationship to other members of the mammalian tree. The discovery of an entire skull of a new gondwanathere genus and species, preserved in three dimensions and described by Krause *et al.*[1] on page 512 of this issue, offers a cornucopia of data not only to solve that mystery but also to reveal further astonishing morphological diversity among early mammals.

The skull, which the authors assign to the species *Vintana sertichi*, was discovered in the Maevarano Formation of Madagascar, which is famous for having produced fossilized birds and non-avian dinosaurs, odd crocodyliforms, a giant frog, fish, turtles and snakes. *Vintana* is dated to the Maastrichtian age of the Upper Cretaceous, between 72 million and 66 million years ago. The authors' detailed computerized tomography scans of the fossil provide a breathtaking look at the cranial anatomy of an animal that combines surprisingly ancient-looking features with others that are advanced specializations (Fig. 1).

Vintana is large — its skull is 12.41 centimetres long and the whole animal is estimated to have weighed almost 9 kilograms. Among



A. JOSEPH GROENKE; B. GARY STAAB

Figure 1 | A gondwanathere skull. The discovery of the skull (**a**) of a gondwanatherian mammal in Madagascar has enabled the first reconstruction (**b**) of one of these enigmatic early mammals. Krause *et al.*[1] estimate that the animal, *Vintana sertichi*, may have been the largest mammal alive 70 million years ago. It had enormous incisors and an expanded cheekbone for the attachment of powerful chewing muscles. It also had large eyes for its size, and a distinctive profile not seen in close relatives.

mammals of the Mesozoic era (which spans the Triassic, Jurassic and Cretaceous periods, from 252 million to 66 million years ago), it is second in size only to the earlier, carnivorous *Repenomamus giganticus* from China[2]. But *Vintana* was a herbivore. Among its specializations are high-crowned molariform teeth with multiple islets of enamel, good for grinding plant matter, and an expanded cheekbone to accommodate the attachment of large chewing muscles; the authors estimate that its bite force was twice that of a rodent of similar size. The animal also had relatively large eyes. This fact, in combination with adaptations of the part of the inner ear concerned with balance, suggests that *Vintana* may have been agile and fast.

Yet the skull also exhibits a few features characteristic of earlier relatives that seem to be the result of evolutionary reversals: the nose, palate and back of the skull contain bones generally thought to have been lost in the mammalian lineage before the common ancestor of living mammals. And its long, down-turned snout would have given it a distinctive appearance. It may be that *Vintana* looks remarkable to us owing to its evolution during Madagascar's long isolation, as Krause and colleagues propose, and indeed tooth wear suggests that it chewed differently from other gondwanatheres. But because we have no comparable cranial material of other gondwanatheres, it may also be that the entire group is similarly unusual.

However outlandish *Vintana* may look, incorporating its skull into phylogenetic analyses at last provides an idea of where gondwanatheres fit within Mammalia. It emerges that they share many characteristics with two groups of Mesozoic mammals now well known from the northern continents — multituberculates and haramiyids — which together comprise a group called Allotheria. *Vintana* is just the latest in an explosion of new discoveries, and controversies, relating to this group. Originally conceived of to unite taxa within the mammalian lineage on the basis of teeth with rows of cusps[3], Allotheria was eventually redefined to include the relatively common multituberculate mammals and the then-mysterious haramiyids, which were known principally from isolated teeth[4].

Because the haramiyids were poorly known, the question of whether the grouping Allotheria reflects an evolutionary relationship has long been debated, so much so that the foremost reference work on Mesozoic mammals[5] has one chapter entitled 'Allotherians' and another denying that haramiyids were true mammals. Last year, two papers describing the first haramiyid skulls and skeletons and including them in new phylogenetic analyses arrived at competing hypotheses: one[6] that haramiyids fell outside Mammalia, and the other[7] that they are closely related to Multituberculata and fell within Mammalia.

This is not a trivial argument — including the earliest known haramiyid fossil in Mammalia pulls mammalian origins back in time to the Triassic period and before the break-up of the supercontinent Pangaea. This scenario favours a model of mammalian diversification during the Jurassic and Cretaceous in which the lineage leading to living marsupial and placental mammals was present for tens of millions of years before diversifying. The alternative model, in which the first mammals would have appeared in the Middle Jurassic, about 175 million years ago, means that living lineages would have diversified more immediately[8]. Last month, three new species of haramiyid were described, from near-complete specimens, with an analysis again favouring the existence of Allotheria and its placement within Mammalia[9]. Now, using a third set of characters, a somewhat different sampling of taxa and all the new information vouchsafed by *Vintana*, Krause and colleagues' model also supports the Allotheria grouping, the placement of *Vintana* within Allotheria, and an ancient origin for Mammalia.

Analyses of fossil taxa with no living descendants, of which Allotheria is an example, produce varying results that are subject to the choice of animals included, the characteristics considered and the type of analysis used. If the inclusion of Allotheria in Mammalia holds up, it may indeed prove to be that Pangaea's fragmentation contributed to mammalian diversification during the Mesozoic. This is not the same as the diversification of the mammalian groups we see today; Mesozoic mammals unrelated to modern marsupials and placentals swam like beavers, soared like flying squirrels and ate like foxes in the henhouse[10]. Living mammals are descended from only a small part of a strange and wonderful early radiation of Mammalia, to which *Vintana* is a particularly informative addition. ■

Anne Weil *is in the Department of Anatomy and Cell Biology, Oklahoma State University Center for Health Sciences, Tulsa, Oklahoma 74107, USA.*
e-mail: anne.weil@okstate.edu

1. Krause, D. W. *et al. Nature* **515,** 512–517 (2014).
2. Hu, Y., Meng, J., Wang, Y. & Li, C. *Nature* **433,** 149–152 (2005).
3. Marsh, O. C. *Am. J. Sci.* **20,** 235–239 (1880).
4. Butler, P. M. *Acta Palaeontol. Pol.* **45,** 317–342 (2000).
5. Kielan-Jaworowska, Z., Cifelli, R. L. & Luo, Z.-X. *Mammals from the Age of Dinosaurs: Origins, Evolution, and Structure* (Columbia Univ. Press, 2004).
6. Zhou, C.-F., Wu, S., Martin, T. & Luo, Z.-X. *Nature* **500,** 163–167 (2013).
7. Zheng, X., Bi, S., Wang, X. & Meng, J. *Nature* **500,** 199–202 (2013).
8. Cifelli, R. L. & Davis, B. M. *Nature* **500,** 160–161 (2013).
9. Bi, S., Wang, Y., Guan, J. & Meng, J. *Nature* http://dx.doi.org/10.1038/nature13718 (2014).
10. Luo, Z.-X. *Nature* **450,** 1011–1019 (2007).

This article was published online on 5 November 2014.

CANCER

Antitumour immunity gets a boost

Five papers extend the list of cancers that respond to therapies that restore antitumour immunity by blocking the PD-1 pathway, and characterize those patients who respond best. SEE LETTERS P.558, P.563, P.568, P.572 & P.577

JEDD D. WOLCHOK & TIMOTHY A. CHAN

The concept that the immune system has a role in controlling cancer is not a recent one. More than a century ago, the surgeon William Coley hypothesized that postoperative bacterial infections could mobilize a patient's own resistance to tumour recurrence, and he developed a mixture of heat-killed bacteria for intratumoral injection that occasionally produced durable regressions[1]. More recently, the elucidation of molecular mechanisms underlying immune regulation has been instrumental in devising strategies to overcome cancer cells' ability to suppress the immune surveillance that would otherwise protect the host from tumour progression[2–4]. One approach to activating these antitumour immune responses has been termed 'checkpoint blockade' — referring to the use of antibodies that block immune-inhibitory pathways switched on by cancer cells. Five papers published in this issue[5–9] reveal a growing list of cancers that respond to checkpoint blockade and describe characteristics of those patients who respond to such therapies.

The immune checkpoints targeted by these therapies serve under normal conditions as molecular brakes, preventing hyperactivity of the T cells of the immune system and, in some cases, preventing autoimmunity[10]. CTLA-4 and PD-1 are two key cell-surface receptors that, when bound by their ligands, trigger such inhibitory pathways and dampen T-cell activity. In the case of the PD-1 pathway, expression of ligands such as PD-L1 on tumour cells can directly lead to the death of T cells expressing

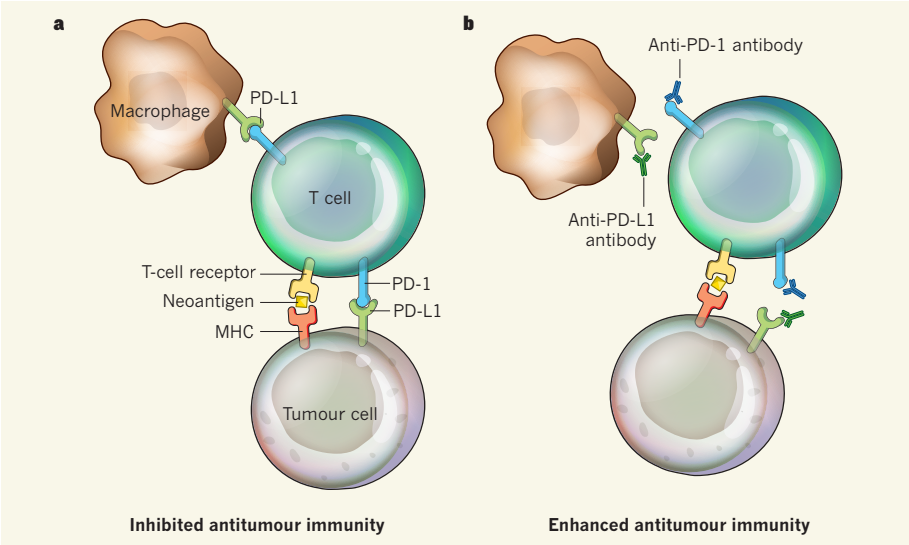


Figure 1 | Checkpoint blockade activates antitumour immunity. **a**, Tumour cells express both cancer-driving mutations and 'passenger' mutations that cause the expression of neoantigens — 'new' molecular structures that, when presented by MHC proteins on the cell surface, are recognized by T cells of the immune system as being foreign, leading to an immune response against the tumour. However, interactions between the receptor PD-1 and its ligand PD-L1, which are expressed on tumour cells, T cells and other immune cells such as macrophages, activate signalling pathways that inhibit T-cell activity and thus inhibit the antitumour immune response. **b**, Antibodies that block the PD-1 pathway by binding to PD-1 or PD-L1 can reactivate T-cell activity and proliferation, leading to enhanced antitumour immunity.

PD-1. Furthermore, engagement of CTLA-4 or PD-1, which are expressed both on T cells and on other immune cells in an inflamed tumour microenvironment, can self-limit the antitumour response. Antibodies that block CTLA-4 (ipilimumab) and PD-1 (pembrolizumab and nivolumab) have been approved to treat patients, and the clinical responses are often durable, with some patients remaining free from disease progression for many years[11–13]. But until recently, little efficacy of these treatments has been noted beyond melanoma and renal-cell carcinoma. Furthermore, the precise cellular events triggered by antibody binding and their exact antigenic targets (the molecular structures to which the antibodies bind) remained unclear.

Powles *et al.*[5] (page 558) and Herbst *et al.*[6] (page 563) present results from a phase I clinical trial of MPDL3280A, a monoclonal antibody that blocks the ligand PD-L1. Herbst and colleagues report that the antibody induces therapeutic responses in patients with non-small-cell lung cancer, melanoma, renal-cell carcinoma and other solid tumours — findings that support the known activity of other antibodies that block the PD-1-pathway in some of these diseases[11,14–16]. Powles and colleagues analyse the effects of this antibody treatment in a larger group of patients with urothelial bladder cancer. Both clinical reports document durable responses in a subset of patients, and that the therapy has low toxicity, with only rare high-grade adverse events. These results substantially expand the spectrum of malignancies in which PD-1-pathway blockade has meaningful clinical activity.

Ever since the earliest reports of the effects of PD-1 blockade[14,15,17,18], PD-L1 expression by tumour cells has been a focus of studies looking for biomarkers that will predict a therapeutic response. Although it is clear that expression of PD-L1 on tumour cells makes it more likely that the patient will respond to PD1-pathway blockade, this is not a binary, static predictive marker. Herbst *et al.* and Tumeh *et al.*[7] (page 568) now reveal that it is not solely tumour-cell expression of PD-L1 that can enrich responses to PD-1-pathway blockade, but that expression of PD-L1 on immune cells infiltrating the tumour is also a key predictor of clinical activity (Fig. 1). Tumeh *et al.* further show, using samples from patients with melanoma that were treated with pembrolizumab, that a certain set of conditions enables PD-1 blockade to mediate tumour regression. These are the presence of $CD8^+$ T cells (a T-cell subset that directly kills its target cells) and immune cells that express PD-1 and PD-L1 at the tumour margin, together with a T-cell population with less-diverse antigen specificity. Taken together, the findings of these two papers suggest that tumours that have already been recognized by the immune system, and so contain infiltrating immune cells bearing PD-1 and PD-L1, are particularly sensitive to immune-checkpoint blockade.

The contributions from Yadav *et al.*[8] (page 572) and Gubin *et al.*[9] (page 577) add another dimension by suggesting that 'passenger' mutations — cancer-cell mutations that do not directly contribute to cancer initiation and progression — play a key part in tumour immunity. Although it is increasingly evident that the new antigens generated by such mutations are targeted by antitumour T cells, identifying which of these neo-antigens are functionally important has been a challenge. Yadav *et al.* sequenced the exomes (the protein-coding regions of the genome) of two mouse tumour-cell lines and compared these with the reference mouse exome to predict candidate neo-antigens in the tumour cells. In parallel, they identified which of the neo-antigens could potentially elicit immune responses by isolating those that bind to major histocompatibility complex (MHC) proteins, which present antigens to T cells, and then analysing the bound peptides by mass spectrometry.

Surprisingly, this process identified only a few candidate neo-antigens, but these were highly immunogenic *in vivo* (that is, they provoked a strong immune response) and were found to be encoded by genes that are unlikely to directly contribute to cancer development, confirming that changes in immunogenicity can result from passenger mutations (Fig. 1). The approach presented in this report is a key advance for the discovery of immunogenic antigens and is applicable to many experimental systems. However, it remains to be seen whether the low numbers of neo-antigens discovered reflects an inherently limited sensitivity of the approach or whether the number of MHC-presented neo-antigens is indeed low.

A previous paper from the research group of Gubin *et al.* showed that a mutant spectrin-β2 protein was responsible for the strong immunogenicity of tumours in a particular mouse model[19]. Now, Gubin and colleagues find that tumours from the same model that become resistant to immune-mediated rejection have lost this neo-antigen. They go on to show that treatment with anti-PD-1 and/or anti-CTLA-4 antibodies enabled the mice to again reject these tumours. Using a similar approach to that of Yadav *et al.*, the authors identify two mutations, in the *Alg8* and *Lama4* genes, that created neo-antigens mediating these effects. Vaccinating mice with these antigens induced tumour rejection at a level comparable to that of checkpoint-blockade therapy, convincingly demonstrating that tumour neo-antigens are potent functional targets of this therapy. This work also corroborates recent findings from another group[20].

These five papers, together with other recent studies, support the hypothesis that immune responses to tumour-specific mutations are central to both natural antitumour immunity and to the antitumour activity generated by checkpoint-blockade therapy. In another twist to the story, a paper[21] just published reports that, in patients with melanoma treated with ipilimumab, specific neo-antigens in the tumour are associated with a favourable clinical response. Intriguingly, these antigens bear a striking similarity to immunogenic antigens derived from bacteria and viruses, suggesting, perhaps, that Coley was on to something. ■

Jedd D. Wolchok *is in the Department of Medicine and the Ludwig Center, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA.* **Timothy A. Chan** *is in the Department of Radiation Oncology and the Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center.*
e-mail: wolchokj@mskcc.org

1. Coley, W. B. *Proc. R. Soc. Med.* **3,** 1–48 (1910).
2. Dunn, G. P., Bruce, A. T., Ikeda, H., Old, L. J. & Schreiber, R. D. *Nature Immunol.* **3,** 991–998 (2002).
3. Koebel, C. M. *et al. Nature* **450,** 903–907 (2007).
4. Shankaran, V. *et al. Nature* **410,** 1107–1111 (2001).
5. Powles, T. *et al. Nature* **515,** 558–562 (2014).
6. Herbst, R. S. *et al. Nature* **515,** 563–567 (2014).
7. Tumeh, P. C. *et al. Nature* **515,** 568–571 (2014).
8. Yadav, M. *et al. Nature* **515,** 572–576 (2014).
9. Gubin, M. M. *et al. Nature* **515,** 577–581 (2014).
10. Page, D. B., Postow, M. A., Callahan, M. K., Allison, J. P. & Wolchok, J. D. *Annu. Rev. Med.* **65,** 185–202 (2014).
11. Robert, C. *et al. Lancet* **384,** 1109–1117 (2014).
12. Topalian, S. L. *et al. J. Clin. Oncol.* **32,** 1020–1030 (2014).
13. Wolchok, J. D. *et al. Ann. Oncol.* **24,** 2174–2180 (2013).
14. Brahmer, J. R. *et al. N. Engl. J. Med.* **366,** 2455–2465 (2012).
15. Topalian, S. L. *et al. N. Engl. J. Med.* **366,** 2443–2454 (2012).
16. Hamid, O. *et al. N. Engl. J. Med.* **369,** 134–144 (2013).
17. Taube, J. M. *et al. Sci. Transl. Med.* **4,** 127ra37 (2012).
18. Brahmer, J. R. *et al. J. Clin. Oncol.* **28,** 3167–3175 (2010).
19. Matsushita, H. *et al. Nature* **482,** 400–404 (2012).
20. Duan, F. *et al. J. Exp. Med.* **211,** 2231–2248 (2014).
21. Snyder, A. *et al. N. Engl. J. Med.* http://dx.doi.org/10.1056/NEJMoa1406498 (2014).

ASTRONOMY

Cosmic triangles and black-hole masses

A geometric measurement of the distance to a nearby galaxy implies a larger mass for its central black hole than previously calculated, and a consequent increase for most other masses of such black holes. SEE LETTER P.528

MARTIN ELVIS

Most distance measurements in astronomy rely on bright sources of light with known power output. How bright an object looks in the sky depends on both its actual light output and its distance from an observer on Earth, so knowing the object's real and apparent brightness allows the distance to be determined. The type 1a supernovae used to discover that the Universe's expansion is accelerating are the best-known examples of such sources. However, a better method for estimating astronomical distances is to use simple geometry, and in this issue Hönig *et al.*[1] (page 528) demonstrate a new geometric-distance method.

If we know the length of the base of an isosceles triangle, with the two long (very long!) sides being the distance to an object, then we can solve for that distance given just one angle. Unfortunately, such a situation is rare in astronomy. The only well-used example is a star's parallax, in which the diameter of Earth's orbit is the base of the triangle and the angle is how much the star moves against the background stars as Earth moves from one side of its orbit around the Sun to the other every six months (Fig. 1). A star that moves 1 arcsecond would be one 'parsec' away. Although the European Space Agency's Gaia satellite is taking parallax to a new level of micro-arcsecond precision, even it cannot use this geometric-distance method on anything distant enough to track the expansion of the Universe.

Quasars, and their less-luminous cousins the active galactic nuclei (AGNs), are objects powered by supermassive black holes that pull (accrete) matter towards them at high speeds. This accretion process releases enough kinetic energy as radiation for some of them to be bright enough to be seen across the Universe. But quasars and AGNs would seem to be unpromising prospects for applying geometric-distance measurements because almost all of the emission from these sources comes from a compact region that cannot be spatially resolved.

However, Hönig *et al.* show that there is a way to obtain the geometric distance to these objects. They invert the triangle used for parallax and put the base of the triangle at the AGN. The sizes of regions in the interior of AGNs are known from a technique called reverberation mapping. The central luminous source in an AGN is tiny, only a few tens of times larger than the supermassive black hole's event horizon — the boundary beyond which no radiation can escape. This source is unstable and varies quite rapidly. Imagine a single flash of light from this small source. This flash travels out at the speed of light. When, after some delay, it encounters gas or dust in its neighbourhood, that material lights up in response; in other words, it 'reverberates'. We can measure how far from the source the lit-up material is just by

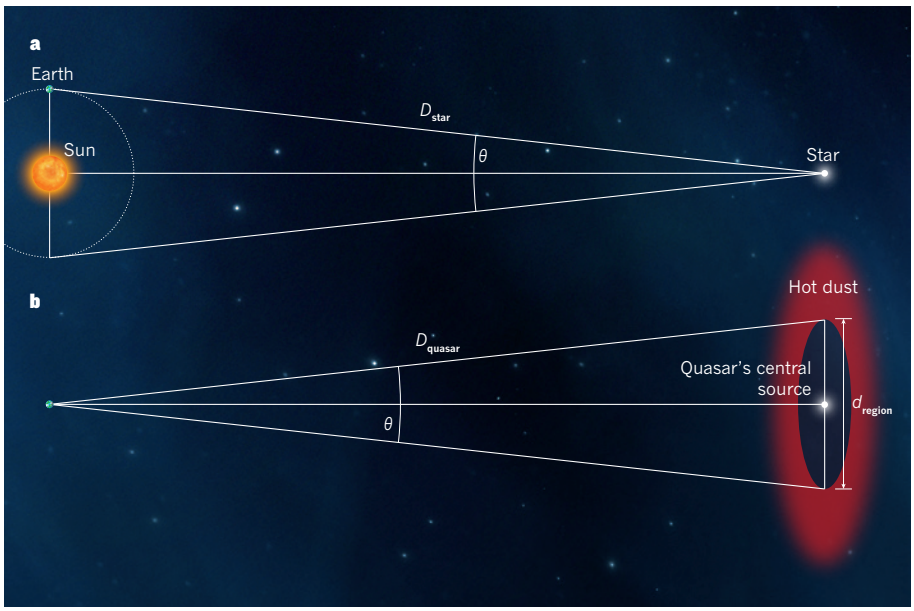


Figure 1 | Geometric distances. **a**, The distance to a star (D_{star}) from Earth can be measured by solving an isosceles triangle. The base of the triangle corresponds to twice the distance between Earth and the Sun, and the angle between its sides (θ) is the angular displacement of the star against a background of distant stars (not shown) as Earth moves from one side of its orbit to the other every six months. **b**, The distance to a quasar or an active galactic nucleus (AGN), D_{quasar}, can be solved if we know the angular size (θ) of some region in the quasar using an interferometer and its linear size (d_{region}) from the time that light takes to move from the quasar's central source out to that region[3]. Hönig *et al.*[1] used the region of hot dust that surrounds the central source of a nearby AGN, and that radiates in the infrared, to measure the distance to this AGN with a roughly 13.5% uncertainty.

multiplying the time delay by the speed of light. This distance gives the length of the base of a triangle (Fig. 1; the actual details of this calculation get complicated[2]). To get the required angle, that is, the angular size of the region that lights up, Margarita Karovska and I suggested[3] using optical interferometers to detect light reverberation from fast-moving clouds located at distances from the central source that are several hundred times the size of the event horizon. But this method has proved to be a step too far for existing interferometers because the angular sizes involved are too small for them to measure.

Hönig and colleagues realized that the hot dust found in AGNs, which radiates in the infrared, was distributed on a sufficiently larger spatial scale than the fast-moving clouds for the Keck interferometer on Mauna Kea, Hawaii, to be able to measure the angular sizes of hot-dust regions in the nearest AGNs at half a milliarcsecond resolution, which is about 100 times better than can be obtained by imaging with the Hubble Space Telescope. The authors relied on the successful Japanese MAGNUM dust-reverberation project[4] for the lengths of the bases of the triangles. Armed with the angular sizes determined using Keck and the lengths from MAGNUM, they have successfully solved the triangle to get the distance to a nearby AGN called NGC 4151, which turns out to be 19.0 (+2.4, −2.6) megaparsecs.

The authors' method is direct, avoiding the usual uncertainty-enhancing steps found in distance measurements that rely on sources with known luminosity. That is the great virtue of using geometry. Although the details of the method are complex, their measurement has an uncertainty of about 13.5% and seems reliable, because the researchers have made several tests of its robustness (see Extended Data for the paper[1]).

This result is noteworthy in two ways. First, NGC 4151 is the AGN with the best-measured black-hole mass from reverberation mapping that also has a black-hole-mass estimate from the kinematics of the surrounding gas and stars[5]. Estimates of black-hole masses from reverberation mapping involve a scale factor that is pinned down by the kinematic mass calculation. Hence, NGC 4151 anchors this reverberation-based, black-hole mass scale, which has been applied to tens of thousands of AGNs and quasars. The kinematic black-hole mass depends on the distance to the AGN, so an accurate distance yields a more reliable scale factor. Hönig and colleagues' distance estimate for NGC 4151 suggests a larger value for this factor and a consequent increase for most black-hole masses of quasars and AGNs obtained using reverberation. Extending this work to more AGNs, as they become better measured in terms of gas and stellar kinematics, will show how much this factor varies from one object to another — a basic requirement if supermassive black-hole masses are to be measured confidently.

Second, and perhaps more importantly, this result is a dramatic demonstration of the power of high-angular resolution in astronomy. Optical and infrared astronomers have mostly plumped for bigger light-collecting areas in the next generation of giant, ground-based telescopes, which will have apertures of 25 to 39 metres in diameter. These telescopes will be equipped with adaptive-optics instruments that correct for the blurring effect of the wobbling of the atmosphere, and will deliver images of higher angular resolution than can Hubble. However, the resolution will still be about three times lower than that of the now-shuttered Keck interferometer.

But Hönig *et al.* have demonstrated that quasar structure and distances can be measured by interferometry. This opens up the prospect of extending AGN size and distance measurements out to the earliest cosmic times, and thus of measuring cosmological properties at distances far beyond where supernovae can take us. New interferometer designs promise great sensitivity increases[6], and the CHARA array has already attained a higher angular resolution than could Keck[7]. Thanks to Hönig *et al.*, we may now have to consider whether some of our resources should soon be put into building a next generation of optical interferometers. ■

Martin Elvis *is at the Harvard-Smithsonian Center for Astrophysics, Cambridge, Massachusetts 02138, USA.*
e-mail: melvis@cfa.harvard.edu

1. Hönig, S. F., Watson, D., Kishimoto, M. & Hjorth, J. *Nature* **515,** 528–530 (2014).
2. Peterson, B. M. & Bentz, M. C. in *Black Holes* (eds Livio, M. & Koekemoer, A. M.) 100–111 (Cambridge Univ. Press, 2011).
3. Elvis, M. & Karovska, M. *Astrophys. J.* **581,** L67–L70 (2002).
4. Suganuma, M. *et al. Astrophys. J.* **639,** 46–63 (2006).
5. Onken, C. A. *et al. Astrophys. J.* **791,** 37 (2014).
6. Buscher, D. F., Creech-Eakman, M., Farris, A., Haniff, C. A. & Young, J. S. *J. Astron. Instrum.* **2,** 1340001 (2013).
7. Pedretti, E., Monnier, J. D., ten Brummelaar, T. & Thureau, N. D. *New Astron. Rev.* **53,** 353–362 (2009).

DEVELOPMENTAL BIOLOGY

Polarize to elongate

An analysis of fruit-fly embryos reveals that receptor proteins of the Toll family direct the oriented cell rearrangements required for the elongation of the head-to-tail axis during development. SEE ARTICLE P.523

ULRICH TEPASS

A central question in developmental biology is how the integrated activities of transcription factors and signalling pathways bring about the cell movements required to create specific tissue shapes. This morphogenesis process forms organs and moulds the animal body. On page 523 of this issue, Paré *et al.*[1] report that Toll-family receptor proteins act as molecular links between the transcriptional machinery that governs head-to-tail patterning and the cellular mechanisms that cause the elongation of this axis in fruit-fly embryos.

During development, the precursor to the trunk region, or germband, of fruit-fly (*Drosophila melanogaster*) embryos elongates about 2.5-fold[2], generating a body with a long head–tail (anterior–posterior) axis and a comparatively narrow back–belly (dorsal–ventral) axis. This elongation, known as convergent extension, is achieved in part by the oriented rearrangement of cells, a process often referred to as neighbour exchange or cell intercalation[3].

Effective cell intercalation is coordinated through the aligned polarization of individual cells in the plane of the tissue. Adherens junctions, which bind cells together, become planar polarized and are disassembled between vertical (anterior–posterior) cell contacts. Subsequently, new horizontal (dorsal–ventral) cell contacts are formed[4,5]. In this way, cells are moved apart by the intercalation of cells from above or below (Fig. 1a). Almost all cells in the germband intercalate in the same orientation, elongating the tissue in the process.

The signalling cascade that controls the subdivision of fly embryos into segments along the anterior–posterior axis includes pair-rule transcription factors. Individual pair-rule genes such as *eve* and *runt* are expressed in alternate segments of the embryo. These genes also govern planar polarity and cell rearrangements during germband extension[2,6]. To investigate how transcriptional regulation by Eve and Runt polarizes cells to direct cell intercalation, Paré and colleagues compared the transcriptional profiles of embryos depleted of Eve or Runt with those of normal embryos. Two of the genes upregulated in Eve- or Runt-depleted embryos encode the Toll receptors Toll-2 and Toll-8. This finding was intriguing because Toll-2 and Toll-8, as well as Toll-6 and Toll-7, are also expressed in different segmental patterns in the *D. melanogaster* germband[7].

Toll receptors are membrane-spanning

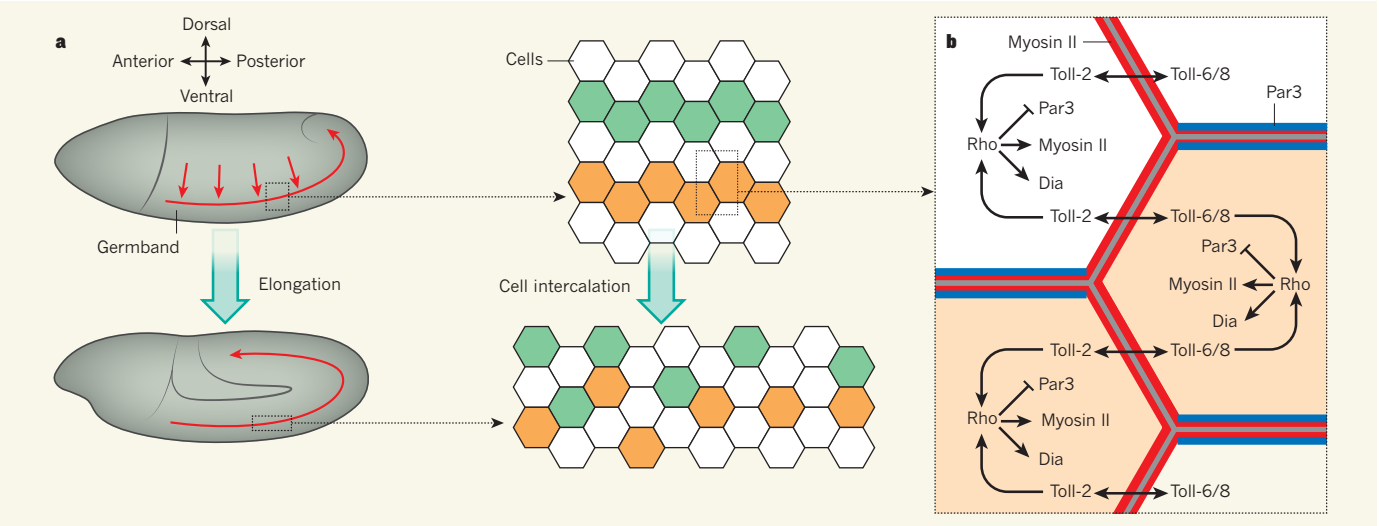


Figure 1 | Cell rearrangement in the *Drosophila melanogaster* germband. **a**, Extension of the precursor to the trunk region, or germband, of fruit-fly embryos involves cell intercalation — cells that lie adjacent to one another along the anterior–posterior axis become separated as cells from above and below (those along the dorsal–ventral axis) move in between them. As a result, the germband elongates in the anterior–posterior direction. Paré *et al.*[1] report that this elongation is mediated by Toll receptor proteins. **b**, A model for the role of Toll receptors in the germband. Interactions between different Toll receptors on adjacent cells may activate the Rho-GTPase signalling pathway (labelled Rho). At vertical contacts, Rho-GTPase signalling promotes the accumulation and activation of the myosin II protein and activation of the Diaphanous (Dia) protein, and it inhibits the Par3 protein, which then dissociates from vertical contacts. Par3 consequently accumulates at horizontal contacts, where it stabilizes adherens junctions.

proteins that are well-established signalling receptors[8]. They can interact with several types of ligand, including other Toll receptors, and have roles in dorsal–ventral patterning, organogenesis and immunity[8]. The authors performed an extensive functional analysis of Toll receptors, which suggested that Toll-2, Toll-6 and Toll-8 cooperate to control planar polarity and cell rearrangement in the germband. Loss of one or two Toll receptors led to local defects in planar polarity, but only the loss of all three receptors caused defects that were similar in strength to the loss of Eve or Runt. At the tissue level, Toll receptors therefore operate redundantly to promote convergent extension of the germband. However, at the cellular level, the receptors are non-redundant because they act on different germband cell populations. Together, these findings implicate Toll receptors in the regulation of planar polarity, and establish them as the long-sought link between transcriptional patterning of the anterior–posterior axis and the cellular mechanisms of axis elongation.

How Eve and Runt regulate the transcription of Toll receptors remains to be explored. The researchers' comparison of the expression patterns of Eve, Runt and Toll receptors and of Toll receptors in Eve- or Runt-depleted embryos suggest that Eve and Runt are not the only factors that regulate the receptors' expression. Other anterior–posterior patterning genes are also required for germband extension[2] and may contribute to this regulation.

How do Toll receptors polarize the cellular machinery that executes cell rearrangement? Paré *et al.* found a first clue in adhesion assays, which showed that Toll-2 in one cell can interact with different receptors (Toll-6 or Toll-8) in adjacent cells, a process known as heterophilic interaction. By contrast, homophilic interactions such as that between Toll-2 and Toll-2 were not observed. This implies that Toll receptors might stimulate their downstream effects through differential enrichment at the two cell surfaces at a vertical interface. However, the distribution of Toll-8 was not obviously polarized, as would be predicted by this model (the authors did not examine Toll-2 and Toll-6 distribution). This suggests either that asymmetries in Toll-receptor localization are subtle or that heterophilic interactions elicit receptor activation without altering receptor distribution. A further concern with this model is that, if heterophilic interactions are essential for receptor activation, one would not expect the receptors' activity to be redundant.

A second clue to how Toll receptors may effect polarity comes from experiments in which Toll-2 and Toll-8 were abnormally expressed. Paré and co-workers report that overexpression of these two receptors in segmental stripes caused the accumulation of myosin II protein at the vertical interfaces of Toll-receptor-expressing and non-expressing cells — a key feature of planar polarity in the germband[3–5]. Although this experiment was performed in late-stage embryos, rather than in the germband, it shows that either Toll-receptor expression boundaries or interfaces of heterophilic receptor interactions can cause the enrichment of myosin II.

Toll-2 can act through the Rho-GTPase signalling pathway, and regulates myosin II during salivary-gland morphogenesis[9]. The Rho pathway also regulates planar polarity and cell rearrangement in the germband. Rho is active at vertical cell contacts, and has several roles in cell intercalation: first, it fosters myosin II accumulation and activation, causing contraction of vertical contacts; second, it activates the protein Diaphanous, which promotes cellular uptake of the cell-adhesion protein DE-cadherin, lowering adhesion at vertical contacts; and third, it causes the Par3 protein to dissociate from vertical contacts and instead to become enriched at horizontal edges, stabilizing adherens junctions at these interfaces[10–12]. Although the authors do not raise this possibility, it is tempting to speculate that asymmetries in receptor distribution at vertical cell contacts — driven by the anterior–posterior patterning machinery — activate Rho signalling to elicit planar polarity (Fig. 1b).

The oriented cell rearrangements that drive convergent extension in tissues other than the fruit-fly germband (for example in many organs with tubes[13–15]) rely on the planar-cell polarity (PCP) signalling pathway for polarization[3]. One possibility is that Toll receptors have functionally replaced the PCP pathway in the germband; this is supported by the fact that the loss of core PCP-pathway components does not interfere with germband extension[6]. However, evidence suggests that the PCP pathway contributes to the planar polarization of myosin II, Par3 and DE-cadherin in this region[16]. It will be interesting to see if and how Toll receptors cooperate with the PCP pathway to polarize germband cells.

Finally, it is important to remember that the cell-intercalation mechanism that is orchestrated by Eve, Runt and Toll receptors seems to account for less than half of germband extension[1,2]. Other mechanisms that have been implicated in germband extension include oriented cell division[17] and large-scale mechanical

forces[18]. Attaining an integrated understanding of morphogenesis remains a formidable challenge. ■

Ulrich Tepass *is in the Department of Cell and Systems Biology, University of Toronto, Toronto, Ontario M5S 3G5, Canada.*
e-mail: u.tepass@utoronto.ca

1. Paré, A. C. *et al. Nature* **515,** 523–527 (2014).
2. Irvine, K. D. & Wieschaus, E. *Development* **120,** 827–841 (1994).
3. Walck-Shannon, E. & Hardin, J. *Nature Rev. Mol. Cell Biol.* **15,** 34–48 (2014).
4. Vichas, A. & Zallen, J. A. *Semin. Cell Dev. Biol.* **22,** 858–864 (2011).
5. Collinet, C. & Lecuit, T. *Prog. Mol. Biol. Transl. Sci.* **116,** 25–47 (2013).
6. Zallen, J. A. & Wieschaus, E. *Dev. Cell* **6,** 343–355 (2004).
7. Kambris, Z., Hoffmann, J. A., Imler, J. L. & Capovilla, M. *Gene Expr. Patterns* **2,** 311–317 (2002).
8. Leulier, F. & Lemaitre, B. *Nature Rev. Genet.* **9,** 165–178 (2008).
9. Kolesnikov, T. & Beckendorf, S. K. *Dev. Biol.* **307,** 53–61 (2007).
10. Levayer, R., Pelissier-Monier, A. & Lecuit, T. *Nature Cell Biol.* **13,** 529–540 (2011).
11. Simões, S. de M. *et al. Dev. Cell* **19,** 377–388 (2010).
12. Simões, S. de M., Mainieri, A. & Zallen, J. A. *J. Cell Biol.* **204,** 575–589 (2014).
13. Nishimura, T., Honda, H. & Takeichi, M. *Cell* **149,** 1084–1097 (2012).
14. Lienkamp, S. S. *et al. Nature Genet.* **44,** 1382–1387 (2012).
15. Affolter, M. & Caussinus, E. *Development* **135,** 2055–2064 (2008).
16. Warrington, S. J., Strutt, H. & Strutt, D. *Development* **140,** 1045–1054 (2013).
17. da Silva, S. M. & Vincent, J. P. *Development* **134,** 3049–3054 (2007).
18. Butler, L. C. *et al. Nature Cell Biol.* **11,** 859–864 (2009).

This article was published online on 2 November 2014.

DIET

Food choices for health and planet

Are you wondering what to prepare for dinner tonight? Data analyses reveal that certain food choices greatly benefit both your health and the environment. But what to do with this evidence remains a challenge to society. SEE ARTICLE P.518

ELKE STEHFEST

Food production has a strong effect on the environment — it is responsible for about 25% of global greenhouse-gas emissions[1], and biodiversity is greatly affected by agricultural land and water use, nutrient loss and fisheries. Within the agricultural sector, livestock farming has the largest environmental footprint[2], and this impact is increasing as traditional diets around the world are being rapidly replaced by diets that are higher in meat, refined sugar and fat. As the scientific basis for the link between diet and the environment grows stronger, the idea has emerged that global dietary changes may contribute to climate-change mitigation[3]. In response, campaigns to promote meat-free days have been launched, such as 'Meatless Monday' in the United States and United Kingdom and 'Veggie Thursday' in Germany and Belgium. In this issue, Tilman and Clark[4] (page 518) show that dietary adjustments would not only reduce greenhouse-gas emissions and agricultural land use, but also greatly reduce individual health risks.

The main novelty of Tilman and Clark's study is that it summarizes strong empirical evidence for the effect of diet on both health and the environment in one publication. For the link between diet and health, the authors compiled data from 18 papers, comprising 8 study cohorts and 10 million person-years of observations, to compare reference diets (including all food groups) to three alternatives: a Mediterranean diet (rich in vegetables, fruit and seafood, but including other foods); a pescetarian diet (including fish and almost no meat); and a vegetarian diet (including dairy products and eggs but almost no meat or fish).

Their review finds a substantial reduction in several negative health indicators for each of these alternative diets compared with the reference diet, including type II diabetes incidence (16–41%, depending on the diet), cancer incidence (7–13%), mortality due to heart disease (20–26%) and overall mortality (0–18%). These effects relate to the fact that the alternative diets contain higher amounts of fruits, vegetables and nuts, and fewer 'empty' calories (energy-containing products that have little other nutritional value, such as alcohol and added sugars) and less meat.

To evaluate the environmental effects of food consumption, the authors assessed two aspects: agricultural greenhouse-gas emissions and changes in land use for agricultural purposes. For agricultural emissions, the authors compiled 120 publications containing 555 life-cycle assessment (LCA) analyses of 82 types of food, which quantify direct emissions along the production chain, including livestock farming, feed production, crop growth, fertilizer application and farm operations. To address land-use change, which has implications for deforestation-associated emissions and biodiversity, they built a simple, transparent model, mostly extrapolating from historical trends. In general, both effects are particularly strong for animal-based food, because livestock, especially ruminants, have a low feed-conversion efficiency, and because ruminants emit methane, a potent greenhouse gas.

Tilman and Clark evaluated both environmental aspects for the year 2050 by comparing a predicted reference diet (based on observed relationships between changes in gross domestic product and food-consumption patterns) with the three alternative diets used for the health analysis (Fig. 1). They find that agricultural emissions would be reduced by 1.2–2.3 gigatonnes of carbon-dioxide carbon equivalents per year (translating to around 30–60% of the projected 2050 emissions from agriculture under the reference diet), and

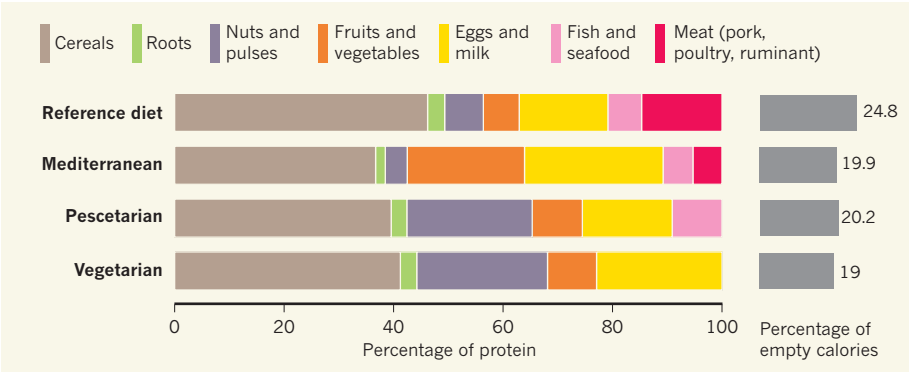


Figure 1 | Dietary transitions. Tilman and Clark[4] find that the global average diet in 2050 ('reference diet'; estimated on the basis of current income-dependent trends in food consumption) will have a high content of 'empty' calories (such as from alcohol and added sugar), meat and other animal products, similar to diets now prevalent in high-income countries. They show that this dietary transition will greatly increase the incidence of non-communicable diseases such as type II diabetes and coronary heart disease, and will be a major contributor to agricultural greenhouse-gas emissions and land clearing. By contrast, their analysis suggests that alternative diets that are richer in fruits, vegetables and pulses could both reduce these environmental effects and offer substantial health benefits.

cropland requirements would be reduced by 450 million to 600 million hectares (about 20–30% of the projected 2050 cropland area for the reference diet) if any one of these alternative diets were adopted by the world's population. The authors conclude that "the implementation of dietary solutions … is a global challenge, and opportunity, of great environmental and public health importance".

How certain are these effects? The link found by Tilman and Clark between diets and health is astonishingly strong, and they used only data that had been corrected for other lifestyle factors. However, as the authors rightly stress, their data are not meant to compare the alternative diets with each other, nor to imply that other diets might not show even higher health benefits. Future research should aim to expand the empirical basis for the connection between diet and health, and to further investigate the mechanisms behind it.

Consistent with other LCA review studies[5], the authors' data analysis shows that greenhouse-gas emissions are highest for ruminant meat, followed by other animal products, and lowest for most cereals, fruits, vegetables and pulses. However, LCA data purely reflect the current state of production systems, and cannot take into account potential efficiency improvements[6]. Other uncertainties arise from the limited scalability of LCA data and agricultural systems in general. For example, it is not clear if vegetable production can be scaled up while maintaining low greenhouse-gas emissions (for example, because there are higher emissions from growth in greenhouses), and changes in livestock consumption and production will also lead to nonlinear effects, feedback and leakages[7] not captured by LCA factors. Marine biologists will also question the scalability of fish production; current levels already overexploit natural stocks, and any increase beyond the sustainable global fisheries catch will have to come from aquaculture, as Tilman and Clark suggest — but expansion of aquaculture will have to rely mostly on land-based feed. However, although such scalability issues should receive further attention, the overall advantage of the alternative diets compared with the reference diet, in terms of emissions, is probably a robust finding.

Predictions of future land requirements for food products and the corresponding environmental consequences are much more uncertain than the emissions predictions, and strongly depend on assumptions about future crop yields, livestock technology and trade. The projected change in global cropland found by Tilman and Clark for the ir 2050 reference diet is higher than the highest estimates in a recent comparison of global agro-economic models[8]. However, the authors find that the reduction in global cropland that would be achieved by the alternative diets is rather constant when varying the most uncertain determinants of their projection in a sensitivity analysis.

With such clear health and environmental benefits of alternative diets, what could be done with this knowledge? First, it can be used by everybody to make informed consumption choices. But individual choices are strongly influenced by the 'food environment' — factors such as shop proximity, food prices, food and nutrition programmes, labelling schemes and community characteristics. Governments and other agencies play a part in shaping these environments to support healthier and more-sustainable food choices, and increased efforts to include both health and environmental factors in dietary guidelines will be key to promoting behavioural change.

Furthermore, addressing consumption should be accompanied by measures on the production side, because regulations at the source of a problem are often the most effective. For example, agriculture and land-use change should be subject to targets and regulations similar to those for the energy and industry sectors. Such interventions will also help to include environmental costs in the price of resource-intensive food products and would therefore further influence individual choices. ■

Elke Stehfest *is in the Department of Climate, Air and Energy, PBL Netherlands Environmental Assessment Agency, Bilthoven, 3720 AH, the Netherlands.*
e-mail: elke.stehfest@pbl.nl

1. Vermeulen, S. J., Campbell, B. M. & Ingram, J. S. I. *Annu. Rev. Environ. Resour.* **37,** 195–222 (2012).
2. Steinfeld, H. *et al. Livestock's Long Shadow: Environmental Issues and Options* (FAO, 2006).
3. Stehfest, E. *et al. Clim. Change* **95,** 83–102 (2009).
4. Tilman, D. & Clark, M. *Nature* **515,** 518–522 (2014).
5. Nijdam, D., Rood, T. & Westhoek, H. *Food Policy* **37,** 760–770 (2012).
6. Havlík, P. *et al. Proc. Natl Acad. Sci. USA* **111,** 3709–3714 (2014).
7. Stehfest, E. *et al. Agric. Syst.* **114,** 38–53 (2013).
8. Von Lampe, M. *et al. Agric. Econ.* **45,** 3–20 (2014).

This article was published online on 12 November 2014.

IMMUNOLOGY

Tolerance lies in the timing

During immune-cell development, potentially self-reactive T cells are eliminated. It emerges that recruitment of a co-receptor bound to the T-cell receptor by the enzyme Lck is the rate-limiting step in this negative selection.

NICHOLAS R. J. GASCOIGNE

Immunological tolerance is a developmental process that enables the immune system to be poised to respond to potential pathogens without inappropriately responding to the body's own molecules. As T cells of the immune system develop in the thymus, those whose T-cell receptor (TCR) binds to a ligand with high strength — an indication that the ligand belongs to a self-molecule — are induced to die. Writing in *Cell*, Stepanek *et al.*[1] identify a molecular mechanism to explain the relationship between TCR recognition strength and this negative selection. The mechanism they propose is based on the 'kinetic proofreading' model originally put forward nearly 20 years ago for T-cell activation[2], which suggests that the induction of an activating signal requires the TCR to bind to its ligand for long enough that the resulting downstream signalling cannot be stopped by the eventual TCR–ligand dissociation. The novelty in Stepanek and colleagues' work is that they have linked the half-life of this interaction to the recruitment of the signalling molecule Lck to the interaction complex, which turns out to be a rather rare event.

Each mature T cell expresses a slightly different TCR that will bind to a complex formed of a specific short peptide (the antigen) and a major histocompatibility complex (MHC) protein on the surface of antigen-presenting cells. The strength of this binding determines the strength of the signal transduced in the T cell. If the antigen is from a foreign organism, this signal needs to be strong enough to activate the cell to respond appropriately to the potential pathogen. But because the TCR repertoire is generated in immature T cells (thymocytes) in a fairly random way, some TCRs will recognize self-antigens, and these cells need to be deleted by negative selection during T-cell development. The challenge in understanding this process has been to determine how a continuous variable — the strength of MHC–peptide binding to the TCR — can be translated into a digital response, in which too-strong signalling leads to death, whereas signalling below this threshold induces positive selection, leading to thymocyte survival and differentiation into mature naive T cells.

In addition to the TCR, developing thymocytes express cell-surface co-receptor proteins, called CD4 and CD8, which bind to MHC class II (MHCII) and class I (MHCI) proteins, respectively. Thymocytes bearing TCRs

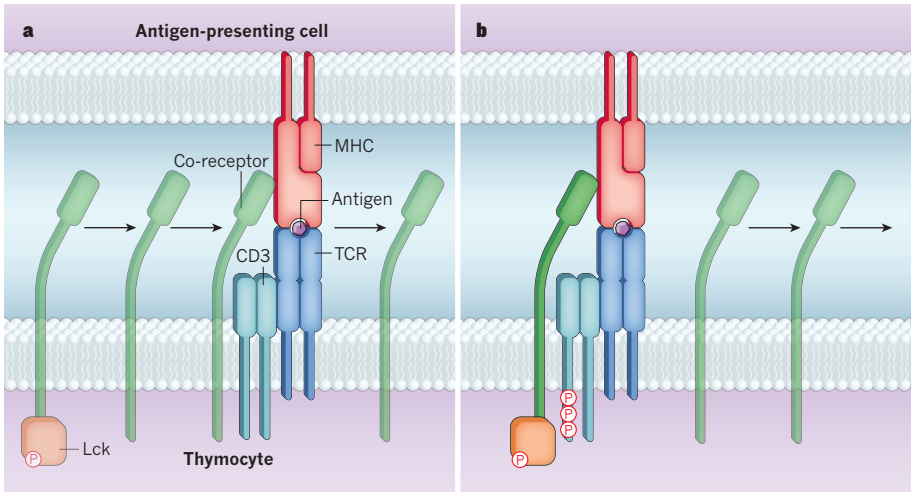


Figure 1 | Dwell time determines selection outcomes in thymocytes. T-cell receptors (TCRs) recognize antigen molecules bound to major histocompatibility (MHC) proteins. The strength of signal transmitted by this interaction determines whether a developing thymocyte is selected to survive (positive selection, induced by not-too-strong signals) or die (negative selection, induced by strong signals indicative of self-antigens). Signal strength is also influenced by the recruitment of T-cell co-receptors to the complex. **a**, Stepanek *et al.*[1] show that co-receptors that are not bound by the protein Lck bind only briefly to MHC before dissociating (indicated by arrows). **b**, By contrast, co-receptors with bound Lck stay associated with the TCR–MHC complex for longer, owing to interactions between the active site of Lck and the CD3 proteins that are associated with the TCR. This longer dwell time leads to the sustained signalling required to induce negative selection. (P denotes protein phosphorylation.)

that recognize MHCII lose CD8 co-receptor expression and become $CD4^+$ T cells, whereas MHCI-restricted thymocytes develop into $CD8^+$ T cells. Expression of these co-receptors enhances cellular sensitivity to antigen, mainly through their recruitment of Lck, a kinase enzyme required for triggering TCR signalling, into the vicinity of the ligated TCR.

The same research group had previously identified ligands that define the threshold between positive and negative selection for MHCI-restricted TCRs[3,4]. Now, they have extended this to MHCII-restricted TCRs, and measured the binding affinity of various MHCII–peptide complexes to the TCR at the negative-selection threshold. They find that the affinity thresholds are similar — around 450 micromolar (μM) for MHCI and around 300 μM for MHCII. However, when they used direct imaging to measure the half-life, or 'dwell time', of these threshold ligands binding to live thymocytes, they found larger differences.

These on-cell binding measurements include the effects of binding of the MHC–peptide complex to both the TCR and the co-receptor; CD8 binds to MHCI more strongly than CD4 binds to MHCII. As a result, the dwell time for the interaction of the MHCI threshold ligands with immature pre-selection thymocytes is about 0.9 seconds, whereas that for MHCII threshold ligands was calculated to be about 0.2 s (unfortunately, it was too short to be measured directly). But a 0.9-s dwell time in the MHCI system corresponds to the dwell time for strongly negative-selecting ligands in the MHCII system, making it difficult to see how the same TCR signal strength can be translated into different functional outcomes in the two systems.

To solve this conundrum, the researchers considered the different roles of the co-receptors in recruiting Lck to the TCR signalling complex. Both CD4 and CD8 bind Lck in a similar way, but CD4 binds rather better than CD8 (ref. 5). Stepanek and co-workers measured the amount of Lck that was bound to the two co-receptors in unstimulated thymocytes, and found that only around 7% of CD4 molecules and around 0.6% of CD8 were bound by Lck. When they factored in the proportion of Lck molecules that were active (on the basis of phosphorylation at a specific site), they were left with a paltry 1.8% and 0.16%, respectively. The authors also showed that a CD8 molecule engineered to use the Lck-binding site from CD4 lowered the negative-selection threshold, such that ligands that were normally at the threshold became strong negative selectors and ligands that were just within the positive-selection range were tipped over the threshold into the negative-selecting realm.

The authors then mathematically modelled the effect of differential co-receptor–Lck coupling on T-cell activation. The model that best fits the data proposes that the probability of a Lck-bound co-receptor being recruited to the TCR–MHC–peptide complex is the crucial factor in kinetic proofreading, because only Lck-bound co-receptors stay bound to the signalling complex long enough to transmit a negative-selection signal (Fig. 1). The probability of CD8–Lck being recruited is lower than that of CD4–Lck, so the TCR complex will need to 'examine' more CD8 molecules than CD4 molecules before it finds one that bears active Lck. Thus, a longer TCR–MHC–peptide dwell time is required when CD8 is involved.

Although this model provides a molecular mechanism for how developing thymocytes translate TCR dwell time into distinct signalling and functional outcomes, and how this varies between MHCI- and MHCII-restricted thymocytes, there are some points that it does not resolve. For example, antigen-independent co-receptor interaction with MHC molecules can concentrate MHC at the interface between an antigen-presenting cell and a thymocyte, with the effect of speeding the rate at which the TCR can bind to MHC–peptide[6]. Because of the higher affinity of CD8 for MHCI than of CD4 for MHCII, this effect will be more marked for MHCI-restricted TCRs and might lower the threshold affinity, but not dwell time, for signalling. Moreover, concentration at this interface also applies to the co-receptor and its associated Lck, so CD8 should be more concentrated than CD4, and the density of CD8-associated Lck at the interface could be higher than estimates obtained from whole-cell analyses.

Another potentially confounding point is that formation of the TCR signalling complex has been identified as a two-step interaction, in which the co-receptor binds to MHC to stabilize the complex only after TCR binding and early signalling events lead to Lck interaction with the TCR complex[7–9]. According to this model, Lck-bound co-receptors preferentially associate with TCRs that have just bound MHC–peptide, which would make the proportion of Lck molecules that are associated with co-receptors less important than in Stepanek and colleagues' model.

Despite such unresolved details, the new model is an attractive variant of the kinetic-proofreading model for T-cell activation, taking into account features of Lck and co-receptor interactions that were not previously accommodated. In particular, co-receptor–Lck interactions change during T-cell differentiation[10], in parallel with changes in antigen sensitivity, and the co-receptor-scanning model provides a simple mechanistic explanation for this phenomenon. ■

Nicholas R. J. Gascoigne *is in the Department of Microbiology, Yong Loo Lin School of Medicine, National University of Singapore, 117597 Singapore.*
e-mail: micnrjg@nus.edu.sg

1. Stepanek, O. *et al. Cell* **159,** 333–345 (2014).
2. McKeithan, T. W. *Proc. Natl Acad. Sci. USA* **92,** 5042–5046 (1995).
3. Daniels, M. A. *et al. Nature* **444,** 724–729 (2006).
4. Naeher, D. *et al. J. Exp. Med.* **204,** 2553–2559 (2007).
5. Wiest, D. L. *et al. J. Exp. Med.* **178,** 1701–1712 (1993).
6. Yachi, P. P., Ampudia, L., Gascoigne, N. R. & Zai, T. *Nature Immunol.* **6,** 785–792 (2005).
7. Jiang, N. *et al. Immunity* **34,** 13–23 (2011).
8. Casas, J. *et al. Nature Commun.* http://dx.doi.org/10.1038/ncomm6624 (2014).
9. Xu, H. & Littman, D. R. *Cell* **74,** 633–643 (1993).
10. Bachmann, M. F. *et al. J. Exp. Med.* **189,** 1521–1530 (1999).

REVIEW

doi:10.1038/nature13855

Belowground biodiversity and ecosystem functioning

Richard D. Bardgett[1] & Wim H. van der Putten[2,3]

Evidence is mounting that the immense diversity of microorganisms and animals that live belowground contributes significantly to shaping aboveground biodiversity and the functioning of terrestrial ecosystems. Our understanding of how this belowground biodiversity is distributed, and how it regulates the structure and functioning of terrestrial ecosystems, is rapidly growing. Evidence also points to soil biodiversity as having a key role in determining the ecological and evolutionary responses of terrestrial ecosystems to current and future environmental change. Here we review recent progress and propose avenues for further research in this field.

The last two decades have witnessed an enormous research effort directed at understanding how biodiversity loss impacts ecosystem functioning, and the influence of this on the goods and services that ecosystems provide[1]. This research has led to the general consensus that biodiversity loss reduces most ecosystem functions and impairs their stability over time, and that functional traits of species have a major role in determining diversity effects[1]. The majority of research on this topic, however, has had an aboveground focus; as a result, our understanding of the functional consequences of biodiversity loss belowground is less well developed. This lack of knowledge hampers our ability to predict the consequences of realistic scenarios of diversity change, especially since belowground biodiversity represents one of the largest reservoirs of biological diversity on Earth[2,3].

Soil communities are extremely complex and diverse, with millions of species and billions of individual organisms being found within a single ecosystem (Table 1), ranging from microscopic bacteria and fungi, through to larger organisms, such as earthworms, ants and moles (Fig. 1). Our understanding of this hidden biodiversity is limited, at least compared to what is known about aboveground diversity. But the last decade has witnessed a growing number of studies testing how belowground communities are distributed in space and time, how they respond to global change and what the consequences of biodiversity change are for plant community dynamics, aboveground trophic interactions, and biogeochemical cycles. Moreover, soil biodiversity research is now entering a new era: awareness is growing among scientists and policy makers of the importance of soil biodiversity for the supply of ecosystem goods and services to human society[4]; and a new generation of tools are available to interrogate the biology of soil and its ecological and evolutionary role.

Here we explore advances in our understanding of the roles of belowground biodiversity, and propose a pathway to further unravel its ecological and evolutionary function in the face of global change. Soil organisms perform a myriad of functions, but we focus here on their roles in nutrient and carbon cycling, plant community dynamics, and the eco-evolutionary responses of ecosystems to global change. We first bring together recent studies that have advanced our understanding of how soil biodiversity is distributed, and how soil diversity regulates ecosystem functions and the structure of terrestrial ecosystems. We then examine how soil biodiversity can mediate impacts of global change on the composition and functioning of terrestrial ecosystems, and explore emerging evidence for the role of soil biodiversity in the evolutionary dynamics of ecosystems. Finally, we highlight research challenges for the new era of soil biodiversity research, and propose a pathway for advancing understanding of the role of soil biodiversity in determining eco-evolutionary responses to global change.

Spatial patterns of soil biodiversity

There is much dispute about how soil biodiversity is distributed across continental and global scales[5,6]. Historically, soil microbial ecologists have been led by the view developed by Baas Becking in 1934 that "everything is everywhere, but, the environment selects". Recent evidence, however, challenges this long-standing view[7]. Studies using molecular techniques, for example, show that bacterial, protistan[8], mycorrhizal[9,10] and faunal[11] taxa in soil have restricted global distributions due to variations in climatic, soil and plant conditions. Also, knowledge that exotic plant species are released from soil-borne pathogens in their new territories challenges the view that everything is everywhere[7], and adds weight to the growing view that most soil organisms are restricted in their global distributions[8–11].

Another long-standing view in ecology is that species richness is maximal in the tropics and gradually declines towards the poles[12]. The global biogeography of soil biota is uncertain due to a lack of data on patterns of occurrence across the world. What data are available indicate that while soil community composition varies across biomes[2,11,13], clear relationships between latitude and species richness do not exist belowground as they do for many taxa aboveground. Other than for termites[14], we are not aware of any evidence that species richness of belowground taxa peaks in the tropics. For soil animals, including nematodes, mites and earthworms[2,15], and mycorrhizal fungi[9,10], the only clear pattern is that diversity is high along most of the latitudinal gradient, and that it drops towards the poles. This suggests a lack of coupling between aboveground and belowground diversity at global scales, a view supported by the finding that areas considered aboveground biodiversity hotspots[16] had lower soil animal diversity than those that are not[11]. This lack of coupling suggests that patterns of aboveground and belowground diversity are governed by different mechanisms[3,12], which are also scale dependent: local soil biodiversity is strongly driven by spatial heterogeneity, and the diversity of microhabitats found within a single, three-dimensional soil profile could be equivalent to that found aboveground within an entire ecosystem[16].

Spatial patterns of soil biodiversity are shaped by a hierarchy of environmental factors, intrinsic population processes, and disturbance and recolonization events operating at different spatial and temporal scales[17]. At the smallest spatial scale (micrometre to millimetre), for example, distribution

[1]Faculty of Life Sciences, Michael Smith Building, The University of Manchester, Manchester M13 9PT, United Kingdom. [2]Department of Terrestrial Ecology, Netherlands Institute of Ecology (NIOO-KNAW), PO Box 50, 6700 AB Wageningen, The Netherlands. [3]Laboratory of Nematology, Wageningen University, PO Box 8123, 6700 ES Wageningen, The Netherlands.

Table 1 | Estimated diversity and abundance of soil taxa according to published literature, supported by expert judgment

Taxon	Diversity per amount soil or area (taxonomic units indicated below)	Abundance (approximate)
Prokaryotes*	100–9,000 cm^{-3}	4–20 × 10^9 cm^{-3}
Fungi†	200–235 g^{-1}‡	100 m g^{-1}
AMF (species)[99–102]	10–20 m^{-2}	81–111 m cm^{-3}
Protists[8]	150–1,200 (0.25 g)$^{-1}$§	10^4–10^7 m^{-2}
Nematodes (genera)[103–105]	10–100 m^{-2}	2–90 × 10^5 m^{-2}
Enchytraeids[106]	1–15 ha^{-1}	12,000–311,000 m^{-2}
Tardigrades[105]‖	?	?
Collembola[105]	20 m^{-2}	1–5 × 10^4 m^{-2}
Mites (Oribatida)[105,107]	100–150 m^{-2}	1–10 × 10^4 m^{-2}
Isopoda[105]	10–100 m^{-2}	10 m^{-2}
Diplopoda[105]	10–2,500 m^{-2}	110 m^{-2}
Earthworms (Oligochaeta)[108]	10–15 ha^{-1}	300 m^{-2}

Units can vary strongly between taxonomic groups, which is in part related to their size, but also whether the organisms are microscopic or macroscopic and whether identifications are based on morphology or on molecular and operational taxonomic units, or whether they are collected per gram, 100 g, or other soil volumes or surfaces. Estimates may differ substantially among soil types and ecosystems. A number of taxa have not been listed, including insects, ectomycorrhizal and ericoid mycorrhizal fungi, and vertebrate organisms (such as moles and voles). Numbers should be taken as preliminary given that most soil species have not yet been described, and because most estimates are based on single ecosystems or regions. We have used data proposed by the following experts: M. van der Heijden (arbuscular mycorrhizal fungi); V. Behan-Pelletier (mites); S. Geisen (protists); H. Helder (nematodes); M. Briones (enchytraeids and tardigrades); P. Lavelle and O. Schmidt (earthworms); supported with published data. Worldwide diversity includes aquatic and marine species[105]. Estimate of 7,000 earthworm species worldwide is gross underestimation due to endemism[109].

* Bacteria and Archaea (genome equivalents)[95]; estimation of worldwide diversity[96].

† Ref. 97 (also includes mycorrhizal fungi)[98].

‡ Operational taxonomic units.

§ Sequences.

‖ There are some 1,500 species of tardigrades known worldwide, but no estimate can be made about numbers of species and numbers of individuals per unit soil.

patterns of soil biota are determined by microscale soil heterogeneity caused by variation in soil architecture and biotic interactions in pore space, including predator–prey interactions, ecosystem engineering by soil animals and rooting patterns of plants. Root exudates also contribute to fine-scale (millimetre to centimetres) spatial patterns in microbial and animal communities[18,19], serving to trigger specific groups of microbes in the rhizosphere, such as nitrate-reducing bacteria and denitrifiers[20] and attract symbiotic organisms to roots, including mycorrhizal fungi and rhizobia[21], entomopathogenic nematodes[22] and microbial antagonists of soil pathogens[23] through chemical signals. At the local scale (centimetres to metres), spatial patterns in soil biota are often explained by variation in the physical and chemical properties of soil, such as soil water, and carbon and nutrient availability, along with the identity of dominant plants, which determines the quantity and quality of substrates entering the soil[3]. At ecosystem, regional and continental scales (metres to thousands of kilometres), other factors such as climate, topography, soil abiotic conditions, such as pH, carbon and nutrient content, and continental isolation, have a more important role[24].

Studying spatial variability of soil biota is challenging given the enormous differences in the size of different soil organisms, which range from 2 μm for bacteria to more than 10 cm for earthworms, and up to a hectare for some soil fungi. Also, while microorganisms and some smaller fauna may be dispersed by wind, dispersal of larger-sized soil biota is limited by active movement, which is generally slow, ranging from 10–100 cm per year for nematodes to tens of metres per year for earthworms. As a result of these factors, coupled with inherent spatial variation in soil abiotic properties and the patchiness of plants, soil organisms are not distributed homogeneously in space; rather, belowground community composition is very fragmentary. In forests, for example, differences in litter quality beneath

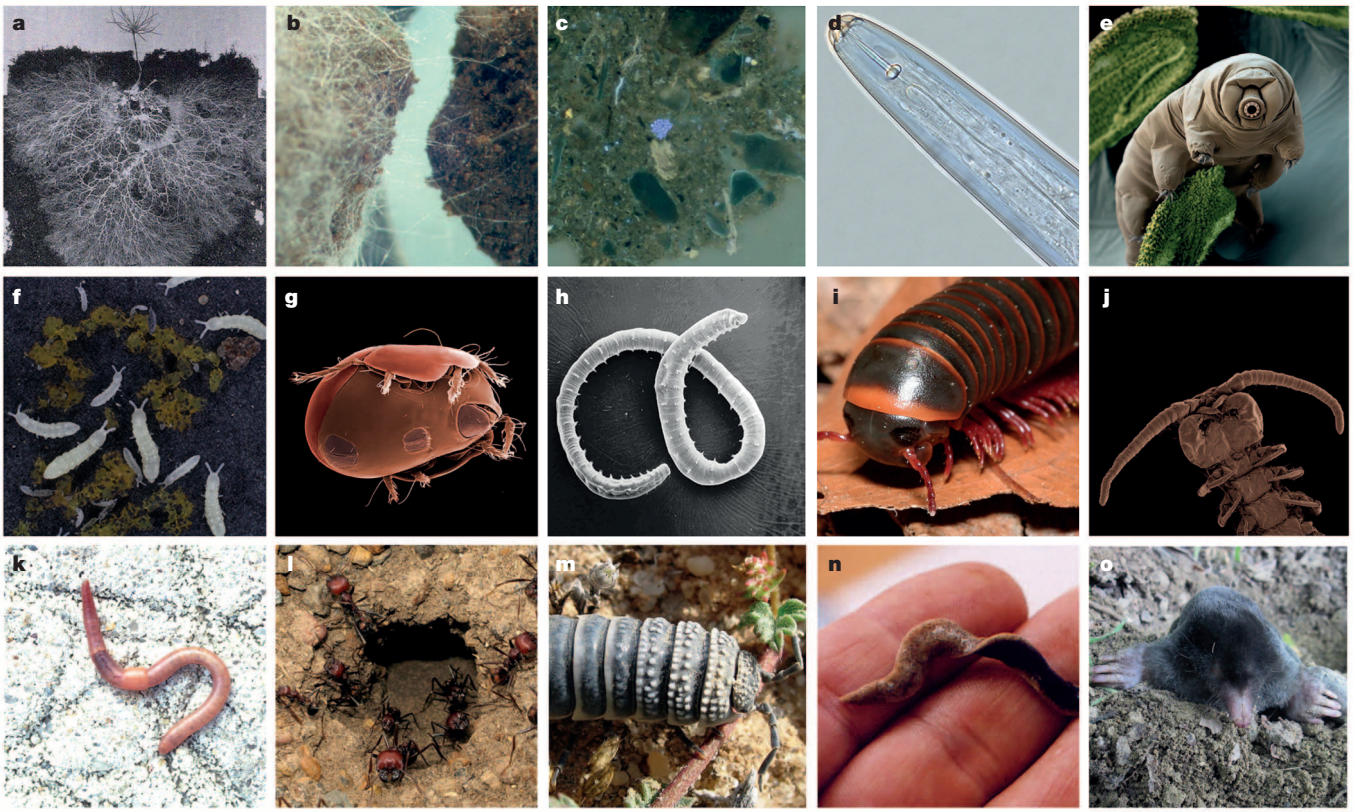


Figure 1 | A selection of organisms of the soil food web. a–o, The selection of organisms includes ectomycorrhizal (**a**) and decomposer fungi (**b**), bacteria (**c**), nematode (**d**), tardigrade (**e**), collembolan (**f**), mite (**g**), enchytraeid worm (**h**), millipede (**i**), centipede (**j**), earthworm (**k**), ants (**l**), woodlice (**m**), flatworm (**n**) and mole (**o**). All photographs are from the European Soil Biodiversity Atlas, courtesy of A. Jones; individual photo credits are: K. Ritz (**b**, **c**); H. van Wijnen (**d**); Water bear in moss, Eye of Science/Science Photo Library (**e**); P. Henning Krog (**f**); D. Walter (**g**); J. Rombke (**h**); J. Mourek (**i**, **j**); D. Cluzeau (**k**); European Soil Biodiversity Atlas, Joint Research Centre (**l**, **n**); S Taiti (**m**); and H. Atter (**o**).

dominant tree species lead to patchy distributions of soil organisms[25], whereas in semi-arid ecosystems, patterning of soil biota and nutrients are related to isolated areas of vegetation that create islands of fertility[26]. Even in cultivated soils, patchy distribution of soil abiotic properties, such as pH and nutrient content, leads to strong patterning of soil organism distribution[27], and in non-vegetated ecosystems, such as the Dry Valleys of Antarctica, distribution patterns of soil organisms are related to spatial patterns in soil carbon and moisture availability[28].

Temporal patterns of soil biodiversity

Surprisingly few studies have examined temporal variability in soil biodiversity, but those that have reveal that population sizes of soil organisms vary over timescales of days to seasons, to decades and millennia. Over short timescales, major drivers of microbial community dynamics are resource pulses, which trigger rapid microbial responses. For example, recent work using molecular tools has shown that sudden increases in soil water availability following rainfall events after prolonged drought cause rapid and sequential resurrection of distinct, phylogenetically clustered groups of microorganisms over timescales ranging from minutes, to hours and days. Moreover, these rapid microbial responses are associated with significant pulses of nitrogen mineralization and CO_2 production from soil[29,30]. Resource pulses from root exudates also drive short-term temporal dynamics of soil biota, with consequences for nutrient cycling and plant nutrient supply. Research has shown, for example, that the time between photosynthesis and the transfer of carbon from leaves to soil organisms is extremely rapid, taking hours in grassland[31] or days in forests[32]. Also, as much as half of this photosynthetic carbon can be lost from soil by respiration within hours or days[19,32], again pointing to the role of root exudation as a major driver of the short-term dynamics of soil communities. Root exudation is also stimulated by biotic interactions with foliar[33] and root herbivores[34], which trigger short-term pulses of microbial activity and nitrogen cycling in the rhizosphere that increase plant nutrient uptake and growth[35,36].

Soil biological communities also vary over seasonal and successional timescales of tens or thousands of years, driven by changes in soil moisture and temperature, and shifts in resource supply in relation to the growth of plants. The few studies that have examined seasonal patterns in soil animal and microbial communities paint a complex picture. Studies of alpine ecosystems, for example, show that microbial communities display a complete turnover between winter and summer, with taxonomically and functionally distinct communities occurring at both times[37]. In agricultural soils, seasonal patterns in soil communities are also highly complex, varying with land use and crop type[38] and from year to year[39]. Even less is known about belowground community development over successional timescales, but a broad pattern appears to exist: at the onset of succession, soil food webs are composed of simple heterotrophic, microbial communities, and photosynthetic and nitrogen-fixing bacteria, but with time they become more complex and stable, with increasing food chain length[40], a reduced role of soil pathogens[41] and greater reliance on mycorrhizal fungi for plant nutrition[42].

Soil biodiversity and ecosystem processes

Ecologists first began to seriously explore the importance of trophic interactions in soil for ecosystem processes in the early 1980s, with microcosm studies revealing their role in stimulating processes of decomposition and nutrient mineralization, and plant nutrient acquisition and growth[43–45]. These studies paved the way for an explosion of research over the last two decades exploring the importance of belowground trophic interactions for example[46,47], and trophic cascades[47,48] for ecosystem processes. Modelling studies have explored the consequences of changes in the architecture and connectedness of soil food webs for energy flux, food web stability and ecosystem processes in real ecosystems[49,50]. While these studies have increased recognition of the functional importance of soil organisms for biogeochemical processes, our understanding of the impact of species loss belowground still has many gaps. From the research that has been done to examine relationships between soil species richness and ecosystem functioning, the main messages to emerge are that for nutrient cycling, diversity effects are of most importance at the low end of the diversity spectrum, and they are dependent, in part, on species traits rather than species richness per se[51–53]. As a result, a common view is that there is high functional redundancy in soil communities for nutrient mineralization, and that changes in belowground community composition, rather than species diversity, are of most importance for ecosystem functioning.

New insights into the functional importance of belowground communities have emerged from studies done in the field. For instance, a field experiment set up across a gradient of sites from the subarctic to the tropics showed that reductions in decomposer functional diversity consistently slowed rates of litter decomposition and carbon and nitrogen cycling[54]. Statistical models have also been used to show that changes in soil food web structure resulting from different land use intensities predicted processes of carbon and nitrogen cycling across a range of European sites, again demonstrating that shifts in soil food webs, in this case due to land use intensification, influence soil functioning under real-world field conditions[55]. In a related study, intensive agriculture was found to impair the resistance and resilience of the soil food web to drought, leading to increased loss of carbon and nitrogen from soil as greenhouse gases and in drainage waters; this was related to a reduction in the 'slow' fungal relative to the 'fast' bacterial energy channel caused by intensive land use[56], suggesting that changes in the asymmetry of these channels, in this case from land use, disrupts ecosystem functioning.

The use of molecular approaches linked to field-based measures of soil carbon cycling has also shown that soil microorganisms regulate impacts of experimental warming on soil carbon and nitrogen dynamics in tallgrass prairie through differential stimulation of microbial populations and the signal intensity of genes involved in decomposition and nitrogen cycling[57]. Similar approaches have been used to reveal the functional role of root-associated fungi involved in ecosystem carbon dynamics in boreal forest[58], and to show how the compositional and functional attributes of soil microbial communities vary across continental gradients[59]. These are just a few examples, and while none explicitly test for soil diversity effects per se, they point at the diverse functional roles of soil organisms in biogeochemical cycles *in situ*.

Soil biodiversity and community dynamics

Over the last two decades, a major focus of soil biodiversity research has been to understand how soil biota impact vegetation dynamics. Traditionally, vegetation dynamics have been explained on the basis of abiotic factors, such as climate and soil physico-chemical properties, and biotic factors such as aboveground herbivory. In recent years, however, it has become widely accepted that vegetation dynamics are also strongly influenced by interactions between plant roots and soil-borne herbivorous, pathogenic, symbiotic and decomposer organisms, especially at local spatial scales[60]. There was already some awareness in the 1990s about the role of soil biota, especially mycorrhizal fungi, root-feeding insects and soil-borne root pathogens as drivers of vegetation dynamics[61–63]. But this became more widely recognized after the turn of the millennium with studies demonstrating the role of plant–soil feedbacks as drivers of plant diversity, abundance and succession[41,64–66], and ecosystem engineers, such as earthworms, in regulating vegetation dynamics[67].

Few studies have tested for effects of soil biodiversity on plant community composition, and these have either focused on soil diversity within single taxonomic groups, such as mycorrhizal fungi[68,69], or on manipulating coarser taxonomic units, for example based on organism body size[70,71]. These studies show that belowground diversity can influence plant community diversity in both positive[68,69,71] and negative[70] ways, which points to the myriad of mechanisms by which complex soil communities impact plant growth, and the potential for differential effects of soil biota to cancel one another out[60]. Indeed, effects of soil biodiversity on vegetation dynamics operate through a variety of biotic interactions, which influence plant performance and vegetation dynamics directly, through altered herbivory, symbiosis, or pathogenesis, or indirectly through changing soil nutrient availability, predation on the plant-feeding organisms or symbionts, or changing interactions between plants and their aboveground multitrophic communities[60,72]. In the short term, these biotic interactions can change

the capacity of plant species to compete, facilitate, and reproduce, whereas longer-term effects influence fitness and evolutionary adaptation.

An area that is especially rich in new discoveries concerns the role of plant secondary metabolites and defence signals in regulating belowground–aboveground interactions[73,74]. It was recently discovered that belowground hyphal networks of arbuscular mycorrhizal fungi act as a conduit for defence signals from plants attacked by herbivorous insects to adjacent non-attacked plants, thereby acting as an early warning system for herbivore attack[75]. Also, foliar and shoot herbivory has been shown to exert a unique soil legacy effect which greatly influences the production of defence chemicals in succeeding plants, and that this legacy effect is mediated by alterations in soil fungal community composition[76]. These studies illustrate that soil biota can impact plant growth by modifying biotic interactions between plants and their natural enemies, but the role of soil biodiversity in these processes remains unresolved.

Considerable recent progress has been made in understanding the role of soil biodiversity in relation to disease suppression and symbiosis, and the use of molecular tools has revealed a previously unexpected diversity of rhizosphere microbes involved[23]. A number of mechanisms have been put forward to explain why and how some soil-borne species contribute to disease suppression, including competition, predation and chemical communication, which collectively contribute to a form of soil biostasis from which many species cannot escape[77]. However, new mechanisms are being proposed, such as the notion that the rhizosphere is a market place where roots and symbionts exchange carbohydrates for nutrients where co-operation can be rewarded, whereas cheating may be discouraged[78].

Eco–evolutionary dynamics and environmental change

Soil biodiversity is currently under threat from a range of anthropogenic pressures, but our understanding of how soil organisms adapt to rapid changes in their environment, whether they can do this fast enough to cope with novel environments, and how this adaptive capacity may relate to the level of soil biodiversity, is limited. A key challenge, therefore, is to determine how soil species respond to rapid environmental change, either through phenotypic plasticity, range shifts or by evolutionary adaptation, how these changes impact aboveground community re-organization and ecosystem functioning, and how the level of soil biodiversity may influence these processes (Fig. 2).

Although scant, evidence is emerging that certain soil organisms have the capacity to respond rapidly to climate change. An analysis of temporal trends in fungal fruiting patterns in southern England, for example, revealed that climate change has advanced the first and extended the last fruiting date of many fungal species, with probable consequences for decomposition processes in soil[79]. Similarly, an analysis of herbarium records in Norway has revealed that the time of fruiting of mushrooms has changed considerably over recent years, although changes differ across taxa[80]. Whether or not these responses were due to plasticity or evolutionary adaptation has not been established. However, it was recently shown that individual species of decomposer fungi can acclimate to climate change, with warm-acclimated fungi reducing their growth and respiration following warming[81]. Given that fungi are the primary agents of decomposition, these results suggest that thermal acclimation of fungi could potentially alter decomposition processes in a warmer world[81]. It was also recently shown that exposure to a new environment can trigger rapid evolutionary change in life history traits of a soil mite, *Sancassania berlesei*, which ultimately alters population dynamics of this species[82]. Although not tested, such eco-evolutionary responses are likely to be widespread with impacts on community dynamics and ecosystem functioning in soil.

Changes in soil microbial community structure also have an impact on evolutionary processes, including patterns of natural selection on plant traits and plant responses to environmental change (Fig. 2). There is a huge body of historic literature reporting how rhizosphere microbes have an impact on plant traits related to nutrient acquisition, drought tolerance, and disease resistance, and ultimately plant fitness, although few studies have been done in non-managed ecosystems[83]. Recent research also shows that modification of soil microbial communities can impact selection on plant traits with, for example, drought-adapted microbial communities increasing plant fitness under this stress[84]. Similar specificity in selective advantage is exemplified by the finding that litter decomposition can be more rapid in soil beneath the host plant species, compared to when beneath a different plant species, the so-called home-field advantage[85]. Home-field advantage effects are not always found and when they are, their strength is highly variable and context dependent. However, recent synthesis suggests that home-field effects are strongest when the quality of 'home' and 'away' litters become more dissimilar, and hence that dissimilarity in plant communities and litter quality between the 'home' and 'away' locations are the most significant drivers of home-field effects[86]. The mechanisms involved in these various community responses still need to be resolved, but it is evident that soil biodiversity has the potential to impact both evolutionary and ecological processes under global change through direct effects

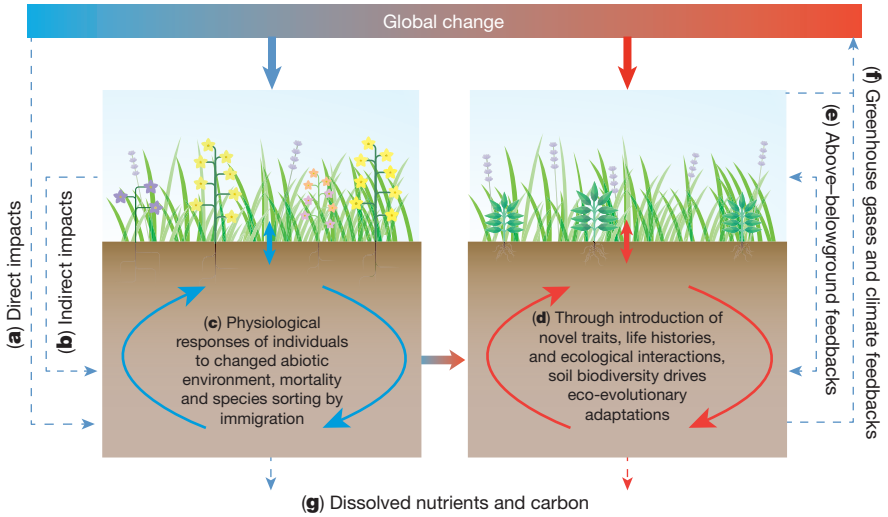


Figure 2 | Belowground responses and feedbacks triggered by climate change. Climate change impacts soil biodiversity directly (**a**), through changes in temperature and moisture, and indirectly (**b**), through shifts in resource supply from plants. Combined, these cause changes in the physiology and growth of individual soil organisms, leading to changes in the diversity and composition of soil communities through altered functional responses and biotic interactions (**c**). As a result, selection for new traits and life histories within soil communities will take place, which in turn drives eco-evolutionary dynamics of aboveground communities (**e**) and ecological feedbacks to ecosystem processes, including greenhouse gas emissions and leaching of dissolved carbon and nutrients from soil (**f**).

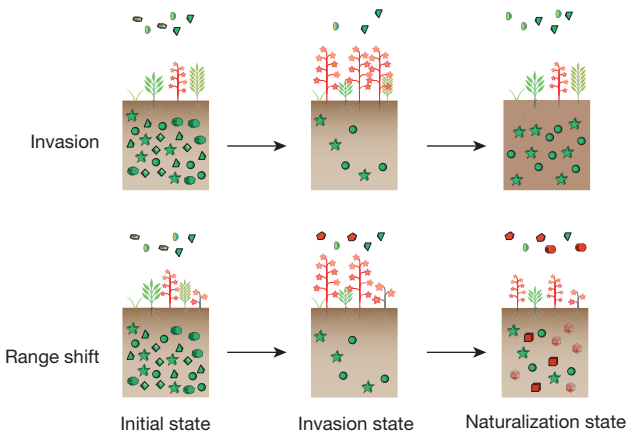
of pathogens, symbionts or root herbivores, as well as by indirect effects involving decomposer organisms in the soil.

Soil biota may also have a role in eco-evolutionary dynamics of introduced exotic plant species (Box 1). An increasing number of studies show that introduced plant species have escaped negative feedback effects from soil biota[87], thus supporting the enemy release hypothesis. As time proceeds following invasion, however, the negative soil feedback may become restored[88,89], which would predict that invasiveness ultimately declines with increasing time since introduction (Box 1 Figure). In one study on *Prunus serotina* (black cherry), soil pathogens were more virulent in the native than in the non-native range[90]. These studies suggest that either the original pathogenic soil biota that cause negative feedback sooner or later are co-introduced, or that native soil pathogens in the new range become more virulent by rapid evolution. Indications of such rapid evolution emerge from studies on the introduced crucifer *Alliaria petiolata* in North America, which showed that soil decomposer communities of recently invaded sites are less capable of decomposing the toxic compounds of the introduced plant than microbial communities from sites that were invaded earlier[91]. Also, following invasion, decomposer richness has been shown to decline and then increase again, suggesting that microbial communities may adapt[92] or become reorganized through dispersal, colonization and establishment processes.

BOX 1

Soil biodiversity, invasions and range shifts

Conceptualized relationship between soil biodiversity and the introduction of invasive exotic plant species that originate from other continents (top row) and plant species that expand their range within the same continent (bottom row). All native biota (plants, aboveground insects and microbes, and belowground microbes and invertebrates) are in green and exotic biota in red. In the initial state (left), abundance of exotic plant species is low and impacts on belowground and aboveground biodiversity are minor. During the invasion state (middle), exotic plant species become disproportionally abundant and might evolve increased competitive ability in the absence of specialized enemies. This process may be more pronounced for intercontinental exotic plant species (top row) than intracontinental range shifts (bottom row), where aboveground enemies also can shift their ranges. In this invasion stage, soil and aboveground biodiversity usually declines owing to loss of specific host plant species, or because of exotic plant species that suppress specific biota by novel chemistry. During the naturalization state (right), soil microbial taxa might rapidly evolve under the influence of the exotic plant species, leading to native pathogens and decomposer organisms adapted to the exotic plant species. As specific pathogenic effects are usually stronger than specific decomposer effects, exotic plants become controlled. If soil biodiversity controls pathogen evolution, the chance of such evolution occurring is greater in the case of intercontinental exotic plant species (top row) than intracontinental range-shifting (bottom row) plant species, as the latter become colonized by range-shifting soil biota from the original range (red symbols). As a result, the abundance of exotic plant species becomes controlled, thereby contributing to their ecological naturalization. Intracontinental range-expanding plant species (bottom row) might enter a naturalization state faster than intercontinental exotics (top row), because aboveground enemies and, later on, soil biota from the native range can shift range as well, but at different rates. Note that these are only some examples for which evidence can be found in current literature; many other scenarios are possible as well, but these need further testing.



Outlook and challenges

The last two decades of soil biodiversity research has revealed that belowground communities are remarkably diverse and that they have a major role in shaping aboveground biodiversity and the functioning of terrestrial ecosystems, as well as their ecological and evolutionary responses to environmental change. One of the biggest challenges for soil ecologists is to integrate this new understanding into existing and novel ecological frameworks in biodiversity-functioning research. Indeed, theory has lagged behind experimental work in soil ecology, and there is a pressing need to adapt existing and develop new theoretical models to explain patterns of belowground community organization and use this to understand their impact on aboveground community dynamics and ecosystem functioning.

There is also a need for improved understanding of the mechanisms that shape complex soil biological communities at different spatial and temporal scales. There is a dearth of information on spatial and temporal patterns of soil biodiversity, and this makes it difficult to develop new models explaining the structuring of soil communities. But the availability of soil biodiversity data is growing rapidly and with this comes the opportunity to develop new frameworks for explaining patterns of community organization at different spatial and temporal scales, and to identify the ecological and evolutionary mechanisms that underlie them. Progress has been made in this area, for example through the use of network analysis to determine patterns of coexistence in soil microbial communities[57,93] and the application of new theories to the stability of food webs in soil[94]. However, a remaining challenge is to merge the complex tangle of biotic interactions that operate in soil into single integrative frameworks that also take into account the structural and chemical complexity of soil.

It is now clear that soil biodiversity affects multiple ecosystem processes, including biogeochemical cycles and eco-evolutionary dynamics in plant and aboveground communities in response to global change. However, questions remain over the relative roles of genetic, species and functional diversity in driving these processes, and the role of extrinsic factors in modulating biodiversity–function relations, such as variations in soil fertility and the structural complexity of soil. Moreover, hardly anything is known about how belowground communities acclimate and adapt to rapid environmental change, although responses of soil biodiversity appear to impact aboveground evolutionary processes, including selection of plant traits in response to environmental change (Box 1 Figure). Such eco-evolutionary responses of belowground communities to rapid environmental change also have the potential to impact community dynamics and ecosystem functioning in soil, but so far this remains unexplored.

Finally, a major goal for soil biodiversity research is to integrate what we learn into sustainable land management decisions, especially regarding new approaches to the maintenance and enhancement of soil fertility for food, feed and biomass production, the prevention of human disease, and the mitigation of climate change. As we highlight here, a new age of research is needed to meet these scientific challenges and to integrate such understanding into future land management and climate change mitigation and adaptation strategies.

Received 29 July; accepted 9 September 2014.

1. Cardinale, B. J. *et al.* Biodiversity loss and its impact on humanity. *Nature* **486,** 59–67 (2012).
2. Decaëns, T. Macroecological patterns in soil communities. *Glob. Ecol. Biogeogr.* **19,** 287–302 (2010).
3. Wardle, D. A. *Communities and Ecosystems: Linking the Aboveground and Belowground Components* (Princeton Univ. Press, 2002).
4. Wall, D. H. *et al.* (eds) *Soil Ecology and Ecosystem Services* (Oxford Univ. Press, 2012).
5. Fierer, N. & Lennon, J. T. The generation and maintenance of diversity in microbial communities. *Am. J. Bot.* **98,** 439–448 (2011).
6. Finlay, B. J. Global dispersal of free-living microbial eukaryote species. *Science* **296,** 1061–1063 (2002).
7. Callaway, R. M. & Maron, J. L. What have exotic plant invasions taught us over the past 20 years? *Trends Ecol. Evol.* **21,** 369–374 (2006).
8. Bates, S. T. *et al.* Global biogeography of highly diverse protistan communities in soil. *ISME J.* **7,** 652–659 (2013).
9. Tedersoo, L. *et al.* Towards global patterns in the diversity and community structure of ectomycorrhizal fungi. *Mol. Ecol.* **21,** 4160–4170 (2012).
10. Öpik, M., Moora, M., Liira, J. & Zobel, M. Composition of root-colonizing arbuscular mycorrhizal fungal communities in different ecosystems around the globe. *J. Ecol.* **94,** 778–790 (2006).
11. Wu, T., Ayres, E., Bardgett, R. D., Wall, D. H. & Garey, J. R. Molecular study of worldwide distribution and diversity of soil animals. *Proc. Natl Acad. Sci. USA* **108,** 17720–17725 (2011).
This study of soils taken from a range of biomes and latitudes showed that cosmopolitan soil animals are extremely rare, and that there is a lack of coupling between aboveground and soil animal diversity at a global scale.
12. Gaston, K. J. Global patterns in biodiversity. *Nature* **405,** 220–227 (2000).
13. Fierer, N., Strickland, M. S., Liptzin, D., Bradford, M. A. & Cleveland, C. C. Global patterns in belowground communities. *Ecol. Lett.* **12,** 1238–1249 (2009).
14. Eggleton, P. & Bignell, D. E. in *Insects in a Changing Environment* (eds Harrington, R. & Stork, N. E.) 473–497 (Academic Press, 1995).
15. Nielsen, U. N. *et al.* Global-scale patterns of soil nematode assemblage structure in relation to climate and ecosystem properties. *Glob. Ecol. Biogeogr.* **23,** 968–978 (2014).
16. Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B. & Kent, J. Biodiversity hotspots for conservation priorities. *Nature* **403,** 853–858 (2000).
17. Ettema, C. H. & Wardle, D. A. Spatial soil ecology. *Trends Ecol. Evol.* **17,** 177–183 (2002).
18. Broeckling, C. D., Broz, A. K., Bergelson, J., Manter, D. K. & Vivanco, J. M. Root exudates regulate soil fungal community composition and diversity. *Appl. Environ. Microbiol.* **74,** 738–744 (2008).
19. Pollierer, M. M., Langel, R., Körner, C., Maraun, M. & Scheu, S. The underestimated importance of belowground carbon input for forest soil animal food webs. *Ecol. Lett.* **10,** 729–736 (2007).
20. Henry, S. *et al.* Disentangling the rhizosphere effect on nitrate reducers and denitrifiers: insight into the role of root exudates. *Environ. Microbiol.* **10,** 3082–3092 (2008).
21. Badri, D. V. & Vivanco, J. M. Regulation and function of root exudates. *Plant Cell Environ.* **32,** 666–681 (2009).
22. Rasmann, S. *et al.* Recruitment of entomopathogenic nematodes by insect-damaged maize roots. *Nature* **434,** 732–737 (2005).
23. Mendes, R. *et al.* Deciphering the rhizosphere microbiome for disease-suppressive bacteria. *Science* **332,** 1097–1100 (2011).
24. Fierer, N. & Jackson, R. B. The diversity and biogeography of soil bacterial communities. *Proc. Natl Acad. Sci. USA* **103,** 626–631 (2006).
This study showed that continental scale patterns of soil bacterial diversity and richness are largely explained by soil pH, diversity and richness being greater in neutral than acidic soils.
25. Saetre, P. & Bååth, E. Spatial variation and patterns of soil microbial community structure in a mixed spruce–birch stand. *Soil Biol. Biochem.* **32,** 909–917 (2000).
26. Delgado-Baquerizo, M., Covelo, F., Maestre, F. T. & Gallardo, A. Biological soil crusts affect small-scale spatial patterns of inorganic N in a semiarid Mediterranean grassland. *J. Arid Environ.* **91,** 147–150 (2013).
27. Robertson, G. P. & Freckman, D. W. The spatial distribution of nematode trophic groups across a cultivated ecosystem. *Ecology* **76,** 1425–1432 (1995).
28. Courtright, E. M., Wall, D. H. & Virginia, R. A. Determining habitat suitability for soil invertebrates in an extreme environment: the McMurdo Dry Valleys, Antarctica. *Antarct. Sci.* **13,** 9–17 (2001).
29. Placella, S. A., Brodie, E. L. & Firestone, M. K. Rainfall-induced carbon dioxide pulses result from sequential resuscitation of phylogenetically clustered microbial groups. *Proc. Natl Acad. Sci. USA* **109,** 10931–10936 (2012).
This study showed that sudden increases in soil water availability following rainfall events after prolonged drought cause rapid and sequential resurrection of distinct, phylogenetically clustered groups of microorganisms, and that these rapid microbial responses are associated with significant pulses of CO_2 production from soil.
30. Placella, S. A. & Firestone, M. K. Transcriptional response of nitrifying communities to wetting of dry soil. *Appl. Environ. Microbiol.* **79,** 3294–3302 (2013).
31. Bahn, M., Schmitt, M., Siegwolf, R., Richter, A. & Brüggemann, N. Does photosynthesis affect grassland soil-respired CO_2 and its carbon isotope composition on a diurnal timescale? *New Phytol.* **182,** 451–460 (2009).
32. Högberg, M. N. *et al.* Quantification of effects of season and nitrogen supply on tree below-ground carbon transfer to ectomycorrhizal fungi and other soil organisms in a boreal pine forest. *New Phytol.* **187,** 485–493 (2010).
33. Hamilton, E. W. & Frank, D. A. Can plants stimulate soil microbes and their own nutrient supply? Evidence from a grazing tolerant grass. *Ecology* **82,** 2397–2402 (2001).
34. Ayres, E., Dromph, K. M., Cook, R., Ostle, N. & Bardgett, R. D. The influence of below-ground herbivory and defoliation of a legume on nitrogen transfer to neighbouring plants. *Funct. Ecol.* **21,** 256–263 (2007).
35. Guitian, R. & Bardgett, R. D. Plant and soil microbial responses to defoliation in temperate semi-natural grassland. *Plant Soil* **220,** 271–277 (2000).
36. Mikola, J. *et al.* Defoliation and patchy nutrient return drive grazing effects on plant and soil properties in a dairy cow pasture. *Ecol. Monogr.* **79,** 221–244 (2009).
37. Schadt, C. W., Martin, A. P., Lipson, D. A. & Schmidt, S. K. Seasonal dynamics of previously unknown fungal lineages in tundra soils. *Science* **301,** 1359–1361 (2003).
38. Lauber, C. L., Ramirez, K. S., Aanderud, Z., Lennon, J. & Fierer, N. Temporal variability in soil microbial communities across land-use types. *ISME J.* **7,** 1641–1650 (2013).
39. Yeates, G. W., Hawke, M. F. & Rijkse, W. C. Changes in soil fauna and soil conditions under *Pinus radiata* agroforestry regimes during a 25-year tree rotation. *Biol. Fertil. Soils* **31,** 391–406 (2000).
40. Neutel, A. M., Heesterbeek, J. A. P. & de Ruiter, P. C. Stability in real food webs: weak links in long loops. *Science* **296,** 1120–1123 (2002).
41. Kardol, P., Bezemer, T. M. & van der Putten, W. H. Temporal variation in plant–soil feedback controls succession. *Ecol. Lett.* **9,** 1080–1088 (2006).
42. Walker, L. R., Wardle, D. A., Bardgett, R. D. & Clarkson, B. D. The use of chronosequences in studies of ecological succession and soil development. *J. Ecol.* **98,** 725–736 (2010).
43. Anderson, J. M., Ineson, P. & Huish, S. A. Nitrogen and cation mobilization by soil fauna feeding on leaf litter and soil organic-matter from deciduous woodlands. *Soil Biol. Biochem.* **15,** 463–467 (1983).
44. Clarholm, M. Interactions of bacteria, protozoa and plants leading to mineralization of soil-nitrogen. *Soil Biol. Biochem.* **17,** 181–187 (1985).
45. Ingham, R. E., Trofymow, J. A., Ingham, E. R. & Coleman, D. C. Interactions of bacteria, fungi, and their nematode grazers - effects on nutrient cycling and plant-growth. *Ecol. Monogr.* **55,** 119–140 (1985).
46. Alphei, J., Bonkowski, M. & Scheu, S. Protozoa, Nematoda and Lumbricidae in the rhizosphere of *Hordelymus europaeus* (Poaceae): Faunal interactions, response of microorganisms and effects on plant growth. *Oecologia* **106,** 111–126 (1996).
47. Laakso, J. & Setälä, H. Sensitivity of primary production to changes in the architecture of belowground food webs. *Oikos* **87,** 57–64 (1999).
48. Hedlund, K. & Öhrn, M. S. Tritrophic interactions in a soil community enhance decomposition rates. *Oikos* **88,** 585–591 (2000).
49. Hunt, H. W. & Wall, D. H. Modelling the effects of loss of soil biodiversity on ecosystem function. *Glob. Change Biol.* **8,** 33–50 (2002).
50. de Ruiter, P. C., Neutel, A. M. & Moore, J. C. Energetics, patterns of interaction strengths, and stability in real ecosystems. *Science* **269,** 1257–1260 (1995).
51. Heemsbergen, D. A. *et al.* Biodiversity effects on soil processes explained by interspecific functional dissimilarity. *Science* **306,** 1019–1020 (2004).
This study showed that functional dissimilarity among detritivorous species, not species number, drives community compositional effects on decomposition and soil respiration.
52. Nielsen, U. N., Ayres, E., Wall, D. H. & Bardgett, R. D. Soil biodiversity and carbon cycling: a review and synthesis of studies examining diversity-function relationships. *Eur. J. Soil Sci.* **62,** 105–116 (2011).
53. Setälä, H., Berg, M. P. & Jones, T. H. in *Biological Diversity and Function in Soils* (eds Bardgett, R. D., Usher, M. B. & Hopkins, D. W.) 236–249 (Cambridge Univ. Press, 2005).
54. Handa, I. T. *et al.* Consequences of biodiversity loss for litter decomposition across biomes. *Nature* **509,** 218–221 (2014).
55. de Vries, F. T. *et al.* Soil food web properties explain ecosystem services across European land use systems. *Proc. Natl Acad. Sci. USA* **110,** 14296–14301 (2013).
This study showed that soil food web properties strongly and consistently predict processes of carbon and nitrogen cycling across land use systems and geographic locations, and they were a better predictor of these processes than agricultural land use intensity.
56. de Vries, F. T. *et al.* Land use alters the resistance and resilience of soil food webs to drought. *Nature Clim. Change* **2,** 276–280 (2012).
57. Zhou, J., Deng, Y., Luo, F., He, Z. L. & Yang, Y. F. Phylogenetic molecular ecological network of soil microbial communities in response to elevated CO_2. *MBio* **2,** e00122–11 (2011).
58. Clemmensen, K. E. *et al.* Roots and associated fungi drive long-term carbon sequestration in boreal forest. *Science* **339,** 1615–1618 (2013).
59. Fierer, N. *et al.* Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc. Natl Acad. Sci. USA* **109,** 21390–21395 (2012).
60. Wardle, D. A. *et al.* Ecological linkages between aboveground and belowground biota. *Science* **304,** 1629–1633 (2004).
61. Bever, J. D., Westover, K. M. & Antonovics, J. Incorporating the soil community into plant population dynamics: the utility of the feedback approach. *J. Ecol.* **85,** 561–573 (1997).

62. Gange, A. C., Brown, V. K. & Sinclair, G. S. Vesicular–arbuscular mycorrhizal fungi: a determinant of plant community structure in early succession. *Funct. Ecol.* **7,** 616–622 (1993).
63. Van der Putten, W. H., van Dijk, C. & Peters, B. A. M. Plant-specific soil-borne diseases contribute to succession in foredune vegetation. *Nature* **362,** 53–56 (1993).
64. Klironomos, J. N. Feedback with soil biota contributes to plant rarity and invasiveness in communities. *Nature* **417,** 67–70 (2002).
65. Maron, J. L., Marler, M., Klironomos, J. N. & Cleveland, C. C. Soil fungal pathogens and the relationship between plant diversity and productivity. *Ecol. Lett.* **14,** 36–41 (2011).
66. Packer, A. & Clay, K. Soil pathogens and spatial patterns of seedling mortality in a temperate tree. *Nature* **404,** 278–281 (2000).
67. Eisenhauer, N. & Scheu, S. Invasibility of experimental grassland communities: the role of earthworms, plant functional group identity and seed size. *Oikos* **117,** 1026–1036 (2008).
68. van der Heijden, M. G. A. *et al.* Mycorrhizal fungal diversity determines plant biodiversity, ecosystem variability and productivity. *Nature* **396,** 69–72 (1998).
69. Wagg, C., Jansa, J., Stadler, M., Schmid, B. & van der Heijden, M. G. A. Mycorrhizal fungal identity and diversity relaxes plant–plant competition. *Ecology* **92,** 1303–1313 (2011).
70. Bradford, M. A. *et al.* Impacts of soil faunal community composition on model grassland ecosystems. *Science* **298,** 615–618 (2002).
71. Wagg, C., Bender, S. F., Widmer, F. & van der Heijden, M. G. A. Soil biodiversity and soil community composition determine ecosystem multifunctionality. *Proc. Natl Acad. Sci. USA* **111,** 5266–5270 (2014).
72. Bezemer, T. M. & van Dam, N. M. Linking aboveground and belowground interactions via induced plant defenses. *Trends Ecol. Evol.* **20,** 617–624 (2005).
73. Biere, A. & Bennett, A. E. Three-way interactions between plants, microbes and insects. *Funct. Ecol.* **27,** 567–573 (2013).
74. Soler, R. *et al.* Root herbivore effects on aboveground multitrophic interactions: patterns, processes and mechanisms. *J. Chem. Ecol.* **38,** 755–767 (2012).
75. Babikova, Z. *et al.* Underground signals carried through common mycelial networks warn neighbouring plants of aphid attack. *Ecol. Lett.* **16,** 835–843 (2013).
This study showed that plants that were not attacked by aboveground aphids induced defence responses when connected by arbuscular mycorrhizal fungi to plants that were attacked by aphids, suggesting that mycorrhizal networks may enable plants to anticipate insect attack by defence induction.
76. Kostenko, O., van de Voorde, T. F. J., Mulder, P. P. J., van der Putten, W. H. & Bezemer, T. M. Legacy effects of aboveground–belowground interactions. *Ecol. Lett.* **15,** 813–821 (2012).
This study showed that feeding on plants by aboveground insects changed soil fungal community composition, which influenced both plant-feeding and carnivorous insects on plants that colonized this soil; indicating that aboveground multitrophic interactions are affected by those of the past through a legacy effect on soil biota.
77. Garbeva, P., Hol, W. H. G., Termorshuizen, A. J., Kowalchuk, G. A. & de Boer, W. Fungistasis and general soil biostasis – a new synthesis. *Soil Biol. Biochem.* **43,** 469–477 (2011).
78. Kiers, E. T. *et al.* Reciprocal rewards stabilize cooperation in the mycorrhizal symbiosis. *Science* **333,** 880–882 (2011).
This study showed that both plants and mycorrhizal fungi have control over mutual interactions and that plants may favour cooperating fungi over cheaters; findings suggest that the rhizosphere is a market place where goods are exchanged by equal partners, rather than where goods are stolen.
79. Gange, A. C., Gange, E. G., Sparks, T. H. & Boddy, L. Rapid and recent changes in fungal fruiting patterns. *Science* **316,** 71 (2007).
80. Kauserud, H. *et al.* Mushroom fruiting and climate change. *Proc. Natl Acad. Sci. USA* **105,** 3811–3814 (2008).
81. Crowther, T. W. & Bradford, M. A. Thermal acclimation in widespread heterotrophic soil microbes. *Ecol. Lett.* **16,** 469–477 (2013).
82. Cameron, T. C., O'Sullivan, D., Reynolds, A., Piertney, S. B. & Benton, T. G. Eco-evolutionary dynamics in response to selection on life-history. *Ecol. Lett.* **16,** 754–763 (2013).
83. Philippot, L., Raaijmakers, J. M., Lemanceau, P. & van der Putten, W. H. Going back to the roots: the microbial ecology of the rhizosphere. *Nature Rev. Microbiol.* **11,** 789–799 (2013).
84. Lau, J. A. & Lennon, J. T. Rapid responses of soil microorganisms improve plant fitness in novel environments. *Proc. Natl Acad. Sci. USA* **109,** 14058–14062 (2012).
This study showed that adaptive plant responses to drought stress are governed by rapid responses of soil microbial communities and suggests that plants may benefit from associations with diverse soil microbial communities when faced with rapid environmental change.
85. Ayres, E. *et al.* Home-field advantage accelerates leaf litter decomposition in forests. *Soil Biol. Biochem.* **41,** 606–610 (2009).
86. Veen, G. F., Freschet, G. T., Ordonez, A. & Wardle, D. A. Litter quality and environmental controls of home-field advantage effects on litter decomposition. *Oikos* http://dx.doi.org/10.1111/oik.01374 (1 July 2014).
87. Gundale, M. J. *et al.* Interactions with soil biota shift from negative to positive when a tree species is moved outside its native range. *New Phytol.* **202,** 415–421 (2014).
88. Diez, J. M. *et al.* Negative soil feedbacks accumulate over time for non-native plant species. *Ecol. Lett.* **13,** 803–809 (2010).
This study showed that non-native plant species introduced longer ago in New Zealand induce more soil pathogenic activity than species introduced more recently, indicating that negative soil feedback toward introduced plant species increases with time since introduction, which may ultimately contribute to their control.
89. Dostál, P., Müllerová, J., Pyšek, P., Pergl, J. & Klinerová, T. The impact of an invasive plant changes over time. *Ecol. Lett.* **16,** 1277–1284 (2013).
90. Reinhart, K. O., Tytgat, T., Van der Putten, W. H. & Clay, K. Virulence of soil-borne pathogens and invasion by *Prunus serotina*. *New Phytol.* **186,** 484–495 (2010).
91. Lankau, R. A., Nuzzo, V., Spyreas, G. & Davis, A. S. Evolutionary limits ameliorate the negative impact of an invasive plant. *Proc. Natl Acad. Sci. USA* **106,** 15362–15367 (2009).
This study showed that introduced exotic plant species produce less phytotoxins with increasing time since introduction, which had strong impacts on soil community functioning; results suggest that effects of invasive species on soil biodiversity may change over time due to evolutionary processes in the plants.
92. Lankau, R. A. Resistance and recovery of soil microbial communities in the face of *Alliaria petiolata* invasions. *New Phytol.* **189,** 536–548 (2011).
93. Barberán, A., Bates, S. T., Casamayor, E. O. & Fierer, N. Using network analysis to explore co-occurrence patterns in soil microbial communities. *ISME J.* **6,** 343–351 (2012).
94. Rooney, N., McCann, K., Gellner, G. & Moore, J. C. Structural asymmetry and the stability of diverse food webs. *Nature* **442,** 265–269 (2006).
95. Torsvik, V., Øvreås, L. & Thingstad, T. F. Prokaryotic diversity–magnitude, dynamics, and controlling factors. *Science* **296,** 1064–1066 (2002).
96. Dykhuizen, D. E. Santa Rosalia revisited: why are there so many species of bacteria? *Antonie van Leeuwenhoek* **73,** 25–33 (1998).
97. Fierer, N. *et al.* Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Appl. Environ. Microbiol.* **73,** 7059–7066 (2007).
98. Taylor, D. L. *et al.* A first comprehensive census of fungi in soil reveals both hyperdiversity and fine-scale niche partitioning. *Ecol. Monogr.* **84,** 3–20 (2014).
99. Öpik, M. *et al.* Global sampling of plant roots expands the described molecular diversity of arbuscular mycorrhizal fungi. *Mycorrhiza* **23,** 411–430 (2013).
100. Kõljalg, U. *et al.* Towards a unified paradigm for sequence-based identification of fungi. *Mol. Ecol.* **22,** 5271–5277 (2013).
101. Kivlin, S. N., Hawkes, C. V. & Treseder, K. K. Global diversity and distribution of arbuscular mycorrhizal fungi. *Soil Biol. Biochem.* **43,** 2294–2303 (2011).
102. Miller, R. M., Reinhardt, D. R. & Jastrow, J. D. External hyphal production of vesicular-arbuscular mycorrhizal fungi in pasture and tallgrass prairie communities. *Oecologia* **103,** 17–23 (1995).
103. Neher, D. A., Wu, J., Barbercheck, M. E. & Anas, O. Ecosystem type affects interpretation of soil nematode community measures. *Appl. Soil Ecol.* **30,** 47–64 (2005).
104. Yeates, G. W. & Bongers, T. Nematode diversity in agroecosystems. *Agric. Ecosyst. Environ.* **74,** 113–135 (1999).
105. Noordijk, J., Kleukers, R. M. J. C., van Nieukerken, E. J. & van Loon, A. J. (eds) *De Nederlandse biodiversiteit – Nederlandse Fauna 10* (Nederlands Centrum voor Biodiversiteit Naturalis & European Invertebrate Survey, 2010).
106. Briones, M. J. I., Ineson, P. & Heinemeyer, A. Predicting potential impacts of climate change on the geographical distribution of enchytraeids: a meta-analysis approach. *Glob. Change Biol.* **13,** 2252–2269 (2007).
107. Norton, R. A. & Behan-Pelletier, V. M. in *A Manual of Acarology* (eds Krantz, G. W. & Walter, D. E.) 430–564 (Texas Tech Univ. Press, 2009).
108. Richard, B. *et al.* Spatial organization of earthworm assemblages in pastures of northwestern France. *Eur. J. Soil Biol.* **53,** 62–69 (2012).
109. Lavelle, P. & Lapied, E. Endangered earthworms of Amazonia: an homage to Gilberto Righi. *Pedobiologia* **47,** 419–427 (2003).

Acknowledgements This work was conceived as part of a symposium on Soil Biodiversity and Ecosystem Functioning at INTECOL, London 2013, which was supported by the British Ecological Society. The work was supported by the European Commission through the project Ecological Function and Biodiversity Indicators in European Soils (EcoFINDERS) (FP7-264465) and an ERC-ADV grant to W.H.v.d.P. We are grateful to P. Brinkman for logistical support, and A. Jones from the Joint Research Centre, Ispra, for providing photographs, and A. Bardgett for compiling Fig. 1.

Author Contributions R.D.B and W.H.v.d.P contributed equally to the planning and writing of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence should be addressed to R.D.B. (richard.bardgett@manchester.ac.uk).

ARTICLE

doi:10.1038/nature13922

First cranial remains of a gondwanatherian mammal reveal remarkable mosaicism

David W. Krause[1], Simone Hoffmann[1], John R. Wible[2], E. Christopher Kirk[3], Julia A. Schultz[4], Wighart von Koenigswald[4], Joseph R. Groenke[1], James B. Rossie[5], Patrick M. O'Connor[6,7], Erik R. Seiffert[1], Elizabeth R. Dumont[8], Waymon L. Holloway[6,7], Raymond R. Rogers[9], Lydia J. Rahantarisoa[10], Addison D. Kemp[3] & Haingoson Andriamialison[11]

Previously known only from isolated teeth and lower jaw fragments recovered from the Cretaceous and Palaeogene of the Southern Hemisphere, the Gondwanatheria constitute the most poorly known of all major mammaliaform radiations. Here we report the discovery of the first skull material of a gondwanatherian, a complete and well-preserved cranium from Upper Cretaceous strata in Madagascar that we assign to a new genus and species. Phylogenetic analysis strongly supports its placement within Gondwanatheria, which are recognized as monophyletic and closely related to multituberculates, an evolutionarily successful clade of Mesozoic mammals known almost exclusively from the Northern Hemisphere. The new taxon is the largest known mammaliaform from the Mesozoic of Gondwana. Its craniofacial anatomy reveals that it was herbivorous, large-eyed and agile, with well-developed high-frequency hearing and a keen sense of smell. The cranium exhibits a mosaic of primitive and derived features, the disparity of which is extreme and probably reflective of a long evolutionary history in geographic isolation.

Mammalia Linnaeus, 1758
Allotheria Marsh, 1880
Gondwanatheria Mones, 1987
Sudamericidae Scillato-Yané and Pascual, 1984
Vintana sertichi gen. et sp. nov.

Etymology. *Vintana* (Malagasy), luck, in reference to the circumstances of discovery of the holotype specimen. Species name after Joseph Sertich, discoverer of UA 9972.

Holotype and only known specimen. A complete and well-preserved cranium, University of Antananarivo (UA) 9972 (Fig. 1; Supplementary Videos 1–3).

Locality and horizon. Locality MAD10-24, Upper Cretaceous (Maastrichtian; 72.1–66.0 Myr ago) Lac Kinkony Member, Maevarano Formation, Mahajanga Basin, northwestern Madagascar[1].

Diagnosis. Taxon differs from all other gondwanatherians in its large size and in exhibiting wear features on molariform tooth crowns indicating a distobuccal (rather than strictly distal) power stroke of the chewing cycle. Full diagnosis in Supplementary Information.

Gondwanatherians are an enigmatic mammalian clade previously represented by only seven valid monotypic genera from the Cretaceous and Palaeogene of South America, Africa, India, Madagascar and the Antarctic Peninsula[2,3]. With the exception of a few dentary fragments, gondwanatherians were previously known only from isolated teeth[3]. No cranial or postcranial material has been assigned to the Gondwanatheria until now, a severe limitation that has left their phylogenetic position within Mammaliaformes uncertain and controversial. Virtually nothing is known of their life habits, aside from inferences that at least the larger sudamericids were herbivorous and ingested an abrasive diet[4–6] and that ferugliotheriids were omnivorous[5]. The cranium described here, from the Late Cretaceous of Madagascar, is remarkably complete and well preserved and only the third known occurrence of a mammaliaform cranium from the Cretaceous of Gondwana[7,8]. It provides an unprecedented opportunity to more reliably assess gondwanatherian relationships and to analyse various aspects of gondwanatherian palaeobiology.

Dental features

UA 9972 represents the first instance in which upper teeth of a gondwanatherian mammal are associated in gnathic material. These teeth provide the foundation for assigning the specimen to a new genus and species and the opportunity to more comprehensively evaluate dental function and diet in gondwanatherians.

Vintana sertichi has an upper dentition consisting of two incisors, no canine, one premolariform tooth, and four molariform teeth in each quadrant. Although the incisors themselves are not preserved, there are two long, curved alveoli in each premaxilla for enlarged, laterally compressed, procumbent, and probably ever-growing incisors that were well separated from the cheek teeth by a long diastema (Figs 1a, c and 2a, b). Based on its alveoli, the single, two-rooted premolariform tooth appears to have been small; neither the left nor right crowns are preserved in UA 9972 (Figs 1c and 2c). Of the eight upper molariform teeth (MF) present in life in *V. sertichi* (four on each side), four heavily worn representatives (left MF2–4, right MF3) are preserved in UA 9972 (Fig. 2c). The molariform cheek teeth have several salient characteristics: large size, hypsodont crowns (extremely worn in UA 9972), quadrangular occlusal profiles, occlusal surfaces worn essentially flat (with heaviest wear in more mesial molariforms, indicating a mesial-to-distal eruption sequence), numerous cementum-filled infundibula, cementum-filled furrows that invaginate from the buccal side but do not extend to the base of the crown, and multiple short roots supporting the periphery of the base of each crown (Fig. 2c, d; Supplementary Videos 4 and 5). The molariform teeth

[1]Department of Anatomical Sciences, Stony Brook University, Stony Brook, New York 11794, USA. [2]Section of Mammals, Carnegie Museum of Natural History, 5800 Baum Boulevard, Pittsburgh, Pennsylvania 15206, USA. [3]Department of Anthropology, University of Texas at Austin, Austin, Texas 78712, USA. [4]Steinmann-Institut für Geologie, Mineralogie und Paläontologie der Universität Bonn, D-53115 Bonn, Germany. [5]Department of Anthropology, Stony Brook University, Stony Brook, New York 11794, USA. [6]Department of Biomedical Sciences, Heritage College of Osteopathic Medicine, Ohio University, Athens, Ohio 45701, USA. [7]Ohio Center for Ecology and Evolutionary Studies, Ohio University, Athens, Ohio 45701 USA. [8]Department of Biology, 221 Morrill Science Center, University of Massachusetts, Amherst, Massachusetts 01003, USA. [9]Geology Department, Macalester College, 1600 Grand Avenue, St Paul, Minnesota 55105, USA. [10]Département de Géologie, Université d'Antananarivo, Antananarivo (101), Madagascar. [11]Département de Paléontologie, Université d'Antananarivo, Antananarivo (101), Madagascar.

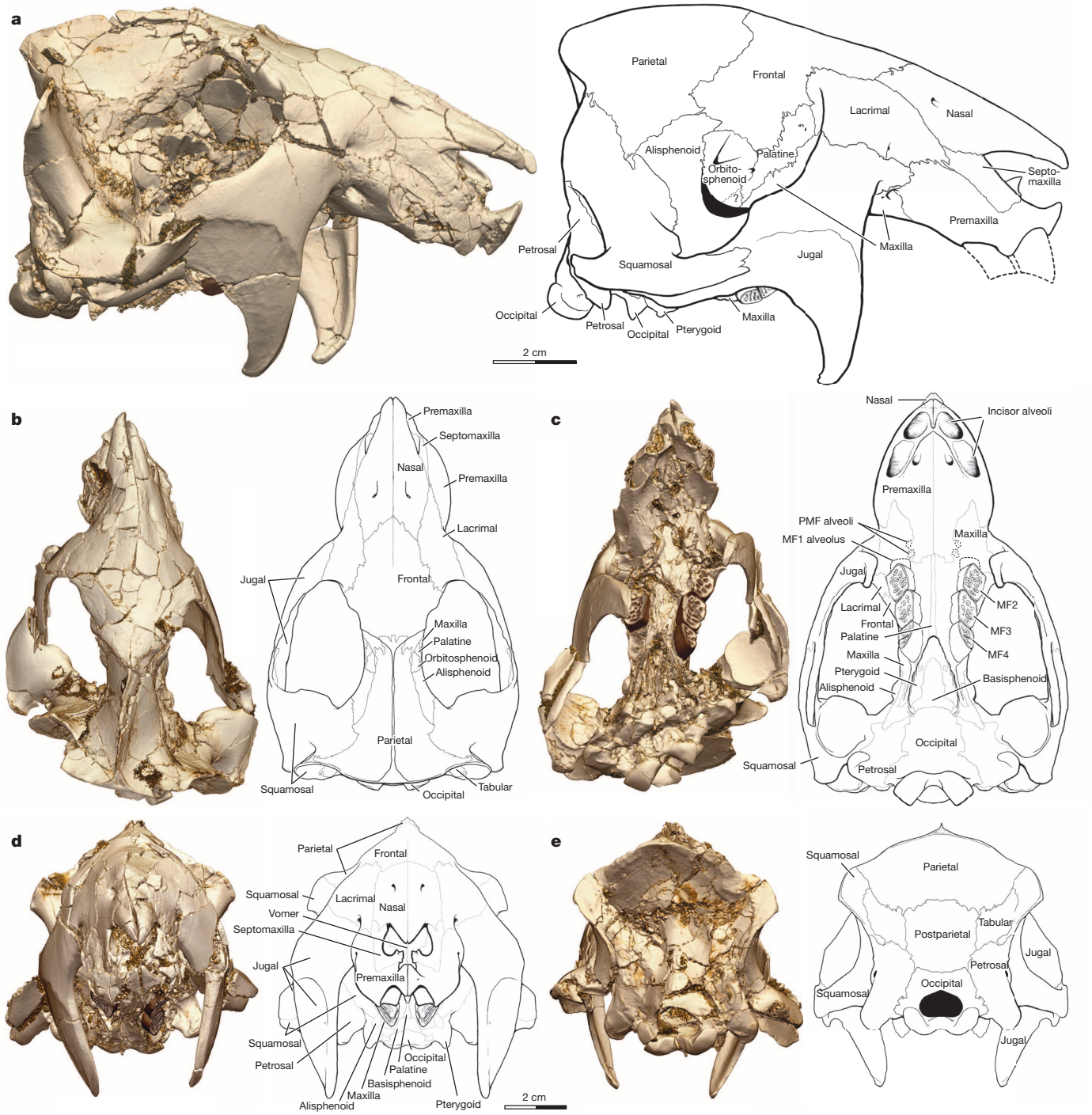


Figure 1 | Cranium of the Cretaceous gondwanatherian mammal *Vintana sertichi*. **a**–**e**, Holotypic specimen, UA 9972, in right lateral (**a**), dorsal (**b**), ventral (**c**), anterior (**d**) and posterior (**e**) views, with micro-computed tomography-based, digitally-rendered image on the left and line drawing reconstruction on the right in each pair. Hypothetical incisor crowns shown in part **a** reconstruction only, and the last three upper molariforms (MF) on both sides and alveoli for the single premolariform tooth (PMF) and MF1 in part **c** reconstruction only.

are also exceptional in that their occlusal surfaces face laterally as much as they do ventrally (Figs 1c, d and 2c).

The enamel microstructure of the molariform teeth of *V. sertichi* retains many of the plesiomorphic characteristics of mammaliaform prismatic enamel (for example, single-layered schmelzmuster; non-decussating, small prisms). However, *V. sertichi* appears to be derived in possessing modified radial enamel consisting of prisms separated by prominent interrow sheets of interprismatic matrix (Extended Data Fig. 1), thus resembling other gondwanatherians from the Late Cretaceous of Madagascar[9,10] and India[9].

Craniofacial features

The well-preserved and complete nature of UA 9972 permits the first insight into the craniofacial morphology of a gondwanatherian mammal. Superficially striking are its short, highly vaulted cranium, large orbits, elongated jugal flanges on widely flaring zygomatic arches, and strong klinorhynchy (Fig. 1; Supplementary Videos 1–3). However, more detailed examination reveals an array of primitive features reminiscent of the most basal mammaliaforms, or even non-mammaliaform cynodonts, coupled with highly derived features unknown in any other Mesozoic mammaliaform.

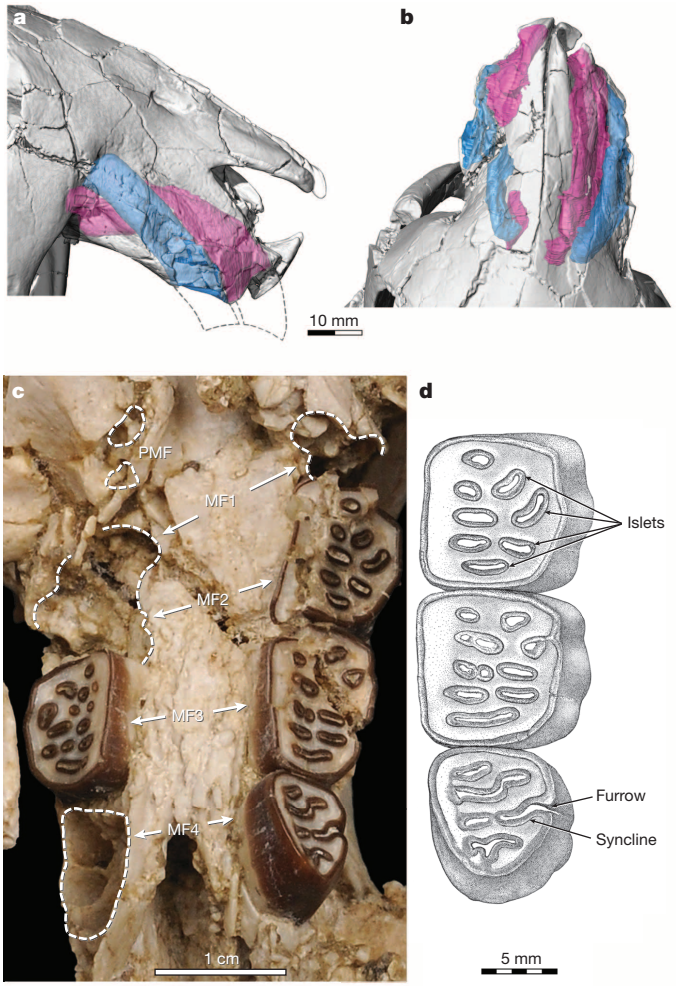


Figure 2 | Upper dentition of *Vintana sertichi*. **a**, **b**, Lateral (**a**) and dorsal (**b**) views of snout region developed from micro-computed tomography data showing position, size and orientation of alveoli of the mesial (red) and distal (blue) incisors. Dashed lines in **a** indicate hypothetical outlines of right incisor crowns. **c**, Ventral (occlusal) view of upper cheek-tooth dentition showing preserved right third molariform (MF3) and left second, third and fourth molariforms (MF2–4) and positions of alveoli for other cheek teeth, including the single right premolariform (PMF). **d**, Occlusal view of reconstructed left MF2–4 showing distribution of cementum-filled islets, furrows and synclines.

The snout of *Vintana* exhibits a number of features that are unique among mammaliaforms (Fig. 1; Extended Data Figs 2a, b; 3a), including the retention of a septomaxilla with both a large posterodorsal facial process and an intranarial process. The lacrimal bone is enormous and extends anteriorly to contact both the septomaxilla and premaxilla. The lacrimal and palatine bones of *Vintana* contribute significantly to the nasal cavity, thereby demonstrating in striking fashion that the identity of adult bones comprising the walls of the cavity varies considerably among mammaliaform taxa. Micro-computed tomography allows the confident identification of a range of nasal cavity structures that have been difficult or impossible to visualize in other Mesozoic mammaliaforms. These include remnants of two turbinal elements, the crista semicircularis and the first ethmoturbinal (including both its lateral and vertical roots), which mark the anterior and posterior boundaries of the lateral recess (fetal pars intermedia). Also clearly identifiable are the basal lamellae of the nasoturbinal and those from which the remaining ethmoturbinals emanated (on the frontal and palatine bones), the peripheral edges of the cribriform plate (on the frontal bone, separating the nasal cavity from the braincase), the posterior transverse lamina (on the palatine and orbitosphenoid bones, separating the cupular recess from the nasopharyngeal canal), and the nasolacrimal canal/groove. Clear evidence of a maxilloturbinal is not preserved. The inferred presence of well-developed turbinal structures in *Vintana* bears witness to the extraordinary conservatism of mammaliaform internal nasal anatomy (despite its unusual bony composition) but is also consistent with the relatively massive size of the olfactory bulbs and the deep, long (~60% of cranial length) nasal cavity.

No Mesozoic mammaliaform has a jugal as enlarged, and with as massive a ventral flange, as that of *Vintana* (Fig. 1). The palate of *Vintana* is very narrow, has a rugose texture, and lacks vacuities or any sizable foramina (other than the incisive foramina anteriorly). Furthermore, the palatines are fused in the midline and extend forward to contact the premaxillae, thus excluding the maxillae from contacting one another in the midline; this feature is unique among Mammaliaformes but reminiscent of the condition in some derived tritylodontids[11].

The lateral and ventral walls of the braincase also reveal a plethora of unusual features (Extended Data Fig. 2c, d), including an alisphenoid that is much larger than the anterior lamina of the petrosal, a robust and compact orbitosphenoid, a large bulbous ectopterygoid process (for possible attachment of the lateral pterygoid muscle) anterior to foramen ovale, the lack of fusion between presphenoid and basisphenoid (despite the advanced age of the individual), a cavum epiptericum that is almost completely floored ventrally, a cavum supracochleare (for the geniculate ganglion of cranial nerve VII) that is separate from the cavum epiptericum (for the semilunar ganglion of cranial nerve V), preservation of a possible process of the ossified pilae antotica and metoptica, and a single foramen (foramen ovale) for the mandibular division of the trigeminal nerve (cranial nerve V_3) between the alisphenoid and the anterior lamina of the petrosal (as in non-mammaliaform cynodonts *Megazostrodon* and *Haldanodon*) but also possibly bordered by the pterygoid (unlike any known Mesozoic mammaliaform).

The basicranium of *Vintana* is also unique among Mesozoic mammaliamorphs in lacking a functional prootic canal and unusual in lacking a channel for the inferior petrosal sinus and any substantial branches of the stapedial system in the middle ear but in possessing an hiatus Fallopii that opens endocranially deep within the posterior wall of the cavum epiptericum (rather than in the lateral trough or at the anterior end of the petrosal), a tympanohyal that abuts the promontorium (as in monotremes), and a fossa incudis that is narrower than the epitympanic recess (as in therians) (Extended Data Figs 2c, 4). *Vintana* is also primitive in retaining a basioccipital lappet on the cochlear housing, as in the mammaliamorphs *Adelobasileus* and *Sinoconodon*. In the occipital region of the cranium, *Vintana* is the only Mesozoic mammaliaform for which both a postparietal and paired tabulars, together comprising the interparietal, have been identified as discrete elements (Fig. 1e; Extended Data Figs 2a, c, 5). However, the presence of these elements may be more commonplace among Mesozoic mammaliaforms, as recently revealed for extant mammals[12]. Finally, many of the cranial elements of *Vintana* contain a remarkable amount of cancellous bone (diploë) compared to those of other Mesozoic mammaliaforms (Extended Data Figs 2a, c and 3a), although a substantial amount is present in the basicranium of the much smaller *Haldanodon*[13].

Endocranial and inner ear features

The mosaicism of derived and primitive features in *Vintana* is also exhibited in its endocranial morphology, a digital reconstruction of which (Extended Data Fig. 3a) reveals that the brain was small and similar in relative size to those of basal mammaliaforms (encephalization quotient = 0.28–0.56). The olfactory bulbs were very large, occupying over 14% of endocranial volume. Unlike the condition in other Mesozoic mammaliaforms, the endocast is strongly flexed (~32° between the olfactory bulbs and the post-olfactory endocast). The cochlear canal, part of the osseous labyrinth of the inner ear, is only slightly curved and short (5.39 mm), only about half the length of the promontorium (Extended Data Fig. 3b). In these regards, *Vintana* resembles various non-mammaliaform cynodonts including tritylodontids, tritheledontids and

Sinoconodon. By contrast, the presence of both primary and secondary osseous laminae, a tractus foraminosus, and Rosenthal's canal represent much more derived characteristics of the mammaliaform inner ear.

Palaeobiology

Vintana sertichi is the largest known mammaliaform from the Mesozoic of Gondwana, superseded in Laurasia only by the eutriconodontan *Repenomamus giganticus* from the Early Cretaceous of China[14]. Estimated from the length of the cranium (124.1 mm long), the body mass of *V. sertichi* was 8.95 kg (95% confidence interval = 5.59–14.32 kg) (see Supplementary Information).

The direction of wear striations, orientation of enamel islets and synclines, and distribution of leading and trailing edges on the molariform teeth of *Vintana* indicate that the direction of the power stroke of the chewing cycle was primarily palinal (distal) (Extended Data Fig. 6), as in haramiyidans[15], multituberculates[16,17] and other gondwanatherians[18,19]. However, *Vintana* appears to be unique among these clades in possessing a significant buccal component to the power stroke. This distobuccal direction is corroborated by biomechanical analyses of the moments generated by the reconstructed primary jaw adductors around the dentary–squamosal joint axis (Extended Data Fig. 7; Supplementary Information). These analyses predict that *Vintana* had significantly higher bite forces than the similarly sized extant rodent *Myocastor* (Supplementary Information). Based on its large size, hypsodont molars and inferred relatively high bite forces, it is likely that *Vintana* had a mixed diet that included large, hard and/or abrasive food items such as roots, seeds or nut-like fruits, not unlike the abrasive, herbivorous diet inferred for other sudamericids[4–6].

When compared with a range of extant mammals, *Vintana* had very large orbits (30–32 mm in diameter) relative to cranial size (Extended Data Fig. 8a). Although orbital size tends to progressively overestimate eye size as body mass and orbital diameter increase[20–22], it is likely that *Vintana* also had relatively large eyes, as is the case for various extant felids, bovids and cervids[23,24] (Extended Data Fig. 8a). Such large eyes could be consistent with either increased sensitivity under low light conditions or enhanced acuity across a range of ambient light levels, depending on eye structure. Furthermore, the radii of curvature of the semicircular canals in *Vintana* are very large (mean = 2.86 mm) for its estimated body mass (Extended Data Fig. 8b). Among living mammals, large semicircular canal radii of curvature are generally associated with large eye size[25]. The semicircular canals of *Vintana* are also almost mutually orthogonal, with angles between the canals ranging from 91° to 94°. This configuration represents only minor deviation from 90° compared to that exhibited by most other Mesozoic synapsids (for example, angles between the anterior and posterior semicircular canals range from 102° to 157° in non-mammaliaform synapsids[26] and 65° to 80° in multituberculates[27], but between 80° and 105° in Cretaceous eutherians[28]). The large size and orthogonality of the semicircular canals suggest that *Vintana* had high vestibular sensitivity to angular head accelerations[29,30]. These vestibular features may have evolved in order to stabilize large eyes during rapid and/or agile locomotion.

The inner ear of *Vintana* also exhibits cochlear primary and secondary osseous laminae and a cochlear canal that is relatively longer than in non-mammaliaform cynodonts but shorter than in extant therians (Extended Data Figs 3b, 8c). These cochlear features are probably associated with the presence of a stiff and short basilar membrane[31], suggesting that *Vintana* had some capacity for high frequency hearing (that is, higher than 20 kilohertz) but that its cochlea may have encoded a more limited range of frequencies than the cochleae of most extant therians.

Finally, *Vintana* had very large olfactory bulbs relative to both its estimated body mass (Extended Data Fig. 8d) and endocranial volume (>14%). In this regard, it resembles some of the most basal mammaliaforms (for example, *Morganucodon*, *Hadrocodium*, *Triconodon*). Because the size of olfactory bulbs likely varies as a function of the number and size of the constituent glomeruli, which receive input from olfactory receptor neurons, it is reasonable to conclude that *Vintana* resembled many extant therian mammals in having an expanded olfactory receptor gene complement and in being able to detect and discriminate among a large number of odorant types[32,33]. These comparative data on the sensory anatomy of *Vintana* indicate that it possessed a distinctive suite of sensory adaptations compared to most other Mesozoic mammaliaforms, including large eyes, some capacity for high frequency hearing, and a keen sense of smell.

Phylogenetic relationships

Gondwanatherians have been variously regarded as Paratheria[34], Xenarthra[35], Multituberculata[6,19], the sister-group to Multituberculata[6,36,37], and Mammalia incertae sedis[2,38]. To assess the relationships of Gondwanatheria to other mammaliaforms, and generic interrelationships within Gondwanatheria, we undertook both parsimony and Bayesian phylogenetic analyses of 87 cynodont taxa (mostly Mesozoic mammaliaforms). This work builds upon previous data sets (see Supplementary Information) by modifying previously used characters, adding several new characters, and scoring several non-therian taxa not incorporated previously (see Supplementary Information for more detailed explanation of data, analysis and results).

Our results indicate that Gondwanatheria are monophyletic, composed primarily of the Sudamericidae, of which *Vintana* is a member, as is the previously unassigned *Greniodon*[3] (Fig. 3). In all analyses Gondwanatheria are placed within the monophyletic Allotheria, including *Haramiyavia*, *Thomasia*, *Arboroharamiya* and Multituberculata (Figs 3, Supplementary Figs 1–4). Relationships among these clades, however, differ between different analytical approaches (see Supplementary Information). The clade containing Gondwanatheria is sister to Multituberculata in the parsimony analysis (Fig. 3, Supplementary Fig. 1), whereas they are nested within the latter in the Bayesian analysis (Supplementary Fig. 2). If this is indeed reflective of the true history of these clades, then several features generally accepted as plesiomorphic within Mammaliamorpha must have re-evolved in the lineage represented by *Vintana* (for example, basioccipital wing overlapping the cochlear housing, large septomaxilla with intranarial process, single trigeminal foramen between anterior lamina and alisphenoid).

Mosaicism and evolution in isolation

Gondwanatherians are a strictly Gondwanan radiation, whereas multituberculates, their closest relatives, are overwhelmingly Laurasian in distribution[2,10,39] (Fig. 3). The early evolution of gondwanatherians remains a mystery that can only be resolved with the discovery of more specimens from the tectonically most active interval of Gondwanan breakup (Middle Jurassic–Early Cretaceous). However, knowledge of the palaeogeographic history of Gondwana provides some insights concerning the lineage to which *Vintana* belongs. Madagascar, together with the Indian subcontinent, separated from Africa approximately 165 million years ago (Myr ago) and became fully isolated from Antarctica and Australia approximately 115–112 Myr ago, with Madagascar and the Indian subcontinent separating from each other about 88 Myr ago[40–42] (Fig. 3, bottom). The basal stock that ultimately led to *Vintana*, at 72–66 Myr ago, was therefore likely isolated on Indo-Madagascar for about 24–27 million years and on Madagascar alone for another approximately 16–22 million years, for a total duration of 40 to almost 50 million years. Interestingly, support for the isolation of the Indo-Malagasy gondwanatherians can be derived from the Bayesian phylogenetic analyses, which identify a node (*Vintana*+*Lavanify*+*Bharattherium*) exclusive of the South American and African forms (see Supplementary Information).

The long period of geographic isolation of Indo-Madagascar and then Madagascar resulted in a latest Cretaceous Malagasy fauna that included a range of other unusual taxa (for example, massive predatory frogs[43], herbivorous crocodyliforms[44], and variously specialized theropod dinosaurs[45–47]). The ghost lineages of these taxa are long and indicate minimum divergence times near the Early–Late Cretaceous boundary (100 Myr ago) or even earlier[40]. Similarly, molecular divergence dates indicate early origins on Madagascar for xenotyphloid blind snakes,

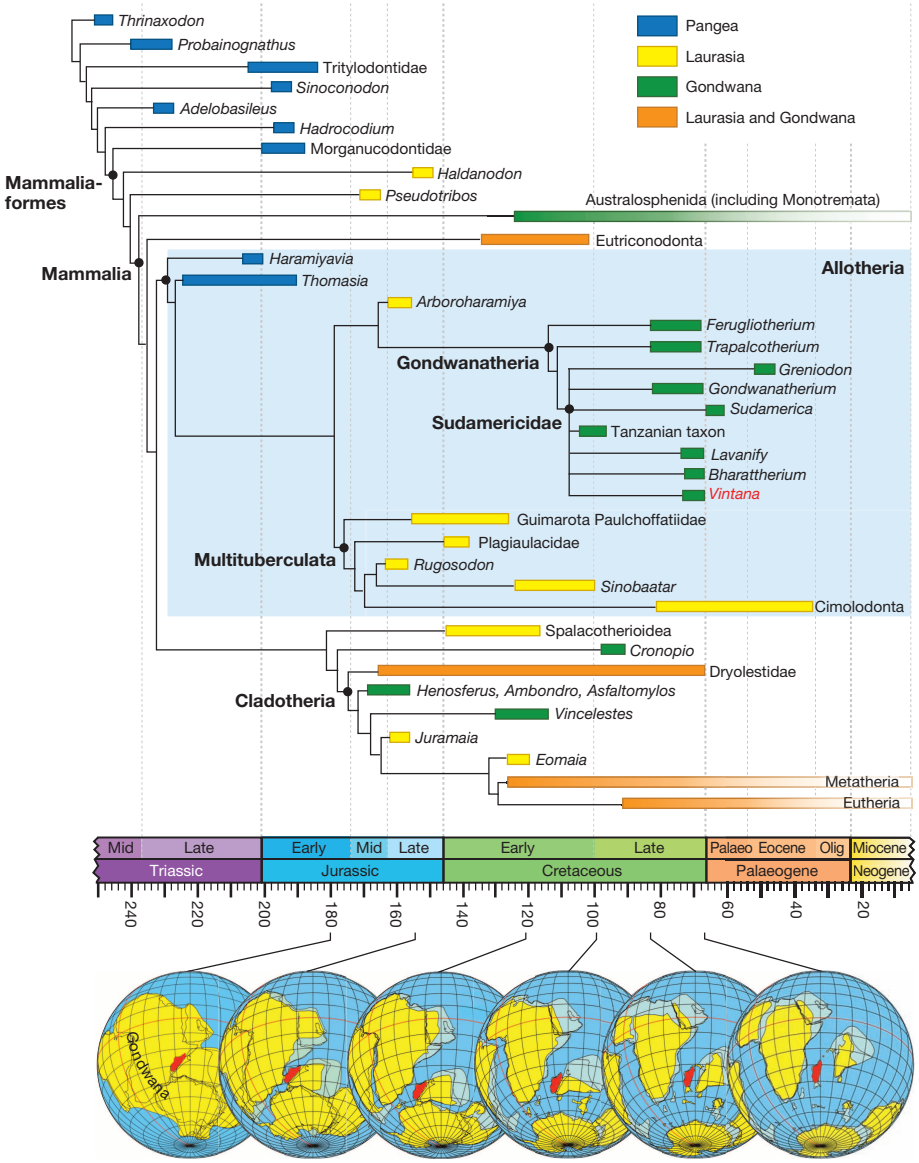


Figure 3 | Relationships of *Vintana* and spatiotemporal distribution of major groups of Mesozoic mammaliaforms and their relatives. Simplified strict consensus tree of 40 equally parsimonious trees (tree length 2,025; consistency index (CI) = 0.305; retention index (RI) = 0.701) derived from analysis of 453 characters and 87 cynodont taxa, with multistate characters unordered and unweighted. Thin black lines represent phylogenetic relationships and thick coloured lines indicate temporal ranges of taxa, with colours indicating supercontinent provenance. *Vintana* is highlighted in red font. Australosphenida includes *Teinolophos*, *Steropodon*, *Bishops*, *Ausktribosphenos* and Monotremata. Allotheria are highlighted in blue. Palaeogeographic maps of Gondwanan supercontinent modified from refs 40 (John Wiley and Sons), 42 and 50 (Elsevier). Madagascar highlighted in red to indicate its progressive geographic isolation through time. See Supplementary Information for taxon and character lists, data matrix, limitations and assumptions, phylogenetic methods, and more detailed explanation of results.

oplurid lizards and podocnemidid turtles[41,48,49]. We hypothesize that these and other lineages were more broadly distributed across Gondwana before fragmentation of the supercontinent, but became isolated on Indo-Madagascar when it separated from first Africa and then Antarctica and Australia. In this spatiotemporal context, *Vintana* appears to have retained several features reflective of its ancestry before early isolation from other Gondwanan landmasses on Indo-Madagascar and then Madagascar alone, but acquired its highly derived, unique morphology during 40–50 million years of evolution in isolation.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 7 July; accepted 7 October 2014.
Published online 5 November 2014.

1. Rogers, R. R. *et al.* A new, richly fossiliferous member comprised of tidal deposits in the Upper Cretaceous Maevarano Formation, northwestern Madagascar. *Cretac. Res.* **44,** 12–29 (2013).
2. Kielan-Jaworowska, Z., Cifelli, R. L. & Luo, Z.-X. *Mammals From the Age of Dinosaurs: Origins, Evolution, and Structure* (Columbia Univ. Press, 2004).
3. Goin, F. J. *et al.* Persistence of a Mesozoic, non-therian mammalian lineage (Gondwanatheria) in the mid-Paleogene of Patagonia. *Naturwissenschaften* **99,** 449–463 (2012).
4. v. Koenigswald, W., Goin, F. & Pascual, R. Hypsodonty and enamel microstructure in the Paleocene gondwanatherian mammal *Sudamerica ameghinoi*. *Acta Palaeontol. Pol.* **44,** 263–300 (1999).
5. Gurovich, Y. Additional specimens of sudamericid (Gondwanatheria) mammals from the early Paleocene of Argentina. *Palaeontology* **51,** 1069–1089 (2008).
6. Gurovich, Y. & Beck, R. The phylogenetic affinities of the enigmatic mammalian clade Gondwanatheria. *J. Mamm. Evol.* **16,** 25–49 (2009).
7. Rougier, G. W. Vincelestes neuquenianus *Bonaparte (Mammalia, Theria) un Primitivo Mamífero del Cretácico Inferior de la Cuenca Neuquina*. PhD thesis, Univ. Buenos Aires (1993).
8. Rougier, G. W., Apesteguía, S. & Gaetano, L. C. Highly specialized mammalian skulls from the Late Cretaceous of South America. *Nature* **479,** 98–102 (2011).
9. Krause, D. W., Prasad, G. V. R., von Koenigswald, W., Sahni, A. & Grine, F. E. Cosmopolitanism among Late Cretaceous Gondwanan mammals. *Nature* **390,** 504–507 (1997).
10. Krause, D. W. Gondwanatheria and ?Multituberculata (Mammalia) from the Late Cretaceous of Madagascar. *Can. J. Earth Sci.* **50,** 324–340 (2013).
11. Clark, J. M. & Hopson, J. A. Distinctive mammal-like reptile from Mexico and its bearing on the phylogeny of the Tritylodontidae. *Nature* **315,** 398–400 (1985).
12. Koyabu, D., Maier, W. & Sánchez-Villagra, M. R. Paleontological and developmental evidence resolve the homology and dual embryonic origin of a mammalian skull bone, the interparietal. *Proc. Natl Acad. Sci. USA* **109,** 14075–14080 (2012).
13. Ruf, I., Luo, Z.-X. & Martin, T. Re-investigation of the basicranium of *Haldanodon exspectatus* (Docodonta, Mammaliaformes). *J. Vertebr. Paleontol.* **33,** 382–400 (2013).
14. Hu, Y., Meng, J., Wang, Y. & Li, C. Large Mesozoic mammals fed on young dinosaurs. *Nature* **433,** 149–152 (2005).
15. Butler, P. M. Review of the early allotherian mammals. *Acta Palaeontol. Pol.* **45,** 317–342 (2000).

16. Krause, D. W. Jaw movement, dental function, and diet in the Paleocene multituberculate *Ptilodus*. *Paleobiology* **8,** 265–281 (1982).
17. Gambaryan, P. P. & Kielan-Jaworowska, Z. Masticatory musculature of Asian taeniolabidoid multituberculate mammals. *Acta Palaeontol. Pol.* **40,** 45–108 (1995).
18. Krause, D. W., Kielan-Jaworowska, Z. & Bonaparte, J. F. *Ferugliotherium* Bonaparte, the first known multituberculate from South America. *J. Vertebr. Paleontol.* **12,** 351–376 (1992).
19. Krause, D. W. & Bonaparte, J. F. Superfamily Gondwanatherioidea: a previously unrecognized radiation of multituberculate mammals in South America. *Proc. Natl Acad. Sci. USA* **90,** 9379–9383 (1993).
20. Schultz, A. H. The size of the orbit and of the eye in primates. *Am. J. Phys. Anthropol.* **26,** 389–408 (1940).
21. Kay, R. F. & Kirk, E. C. Osteological evidence for the evolution of activity pattern and visual acuity in primates. *Am. J. Phys. Anthropol.* **113,** 235–262 (2000).
22. Kirk, E. C. Effects of activity pattern on eye size and orbital aperture size in primates. *J. Hum. Evol.* **51,** 159–170 (2006).
23. Ritland, S. M. *The Allometry of the Vertebrate Eye*. PhD thesis, Univ. Chicago (1982).
24. Ross, C. F. & Kirk, E. C. Evolution of eye size and shape in primates. *J. Hum. Evol.* **52,** 294–313 (2007).
25. Kemp, A. D. & Kirk, E. C. Eye size and visual acuity influence vestibular anatomy in mammals. *Anat. Rec.* **297,** 781–790 (2014).
26. Olson, E. C. Origin of mammals based upon the cranial morphology of therapsid suborders. *Spec. Pap. Geol. Soc. Am.* **55,** 1–130 (1944).
27. Hurum, J. H. The inner ear of two Late Cretaceous multituberculate mammals, and its implications for multituberculate hearing. *J. Mamm. Evol.* **5,** 65–93 (1998).
28. Ekdale, E. G. Comparative anatomy of the bony labyrinth (inner ear) of placental mammals. *PLoS ONE* **8,** e66624 (2013).
29. Yang, A. & Hullar, T. E. Relationship of semicircular canal size to vestibular-nerve afferent sensitivity in mammals. *J. Neurophysiol.* **98,** 3197–3205 (2007).
30. Berlin, J. C., Kirk, E. C. & Rowe, T. B. Functional implications of ubiquitous semicircular canal non-orthogonality in mammals. *PLoS ONE* **8,** e79585 (2013).
31. Luo, Z. X., Ruf, I. & Martin, T. The petrosal and inner ear of the Late Jurassic cladotherian mammal *Dryolestes leiriensis* and implications for ear evolution in therian mammals. *Zool. J. Linn. Soc.* **166,** 433–463 (2012).
32. Mori, K., Nagao, H. & Yoshihara, Y. The olfactory bulb: coding and processing of odor molecule information. *Science* **286,** 711–715 (1999).
33. Niimura, Y. Olfactory receptor multigene family in vertebrates: from the viewpoint of evolutionary genomics. *Curr. Genomics* **13,** 103–114 (2012).
34. Scillato-Yané, G. J. & Pascual, R. Un peculiar Paratheria, Edentata (Mammalia) del Paleoceno de Patagonia. *Primeras J. Argent. Paleontol. Vertebr.*, abstr. 15 (1984).
35. Scillato-Yané, G. J. & Pascual, R. Un peculiar Xenarthra del Paleoceno medio de Patagonia (Argentina). Su importancia en la sistemática de los Paratheria. *Ameghiniana* **21,** 173–176 (1985).
36. Gurovich, Y. *Bio-evolutionary Aspects of Mesozoic Mammals: Description, Phylogenetic Relationships and Evolution of the Gondwanatheria (Late Cretaceous and Paleocene of Gondwana)*. PhD thesis, Univ. Buenos Aires (2006).
37. Pascual, R. & Ortiz-Jaureguizar, E. The Gondwanan and South American episodes: two major and unrelated moments in the history of the South American mammals. *J. Mamm. Evol.* **14,** 75–137 (2007).
38. Pascual, R., Goin, F. J., Krause, D. W., Ortiz-Jaureguizar, E. & Carlini, A. A. The first gnathic remains of *Sudamerica*: implications for gondwanathere relationships. *J. Vertebr. Paleontol.* **19,** 373–382 (1999).
39. Rich, T. H. *et al.* An Australian multituberculate and its palaeobiogeographic implications. *Acta Palaeontol. Pol.* **54,** 1–6 (2009).
40. Ali, J. R. & Krause, D. W. Late Cretaceous bioconnections between Indo-Madagascar and Antarctica: refutation of the Gunnerus Ridge causeway hypothesis. *J. Biogeogr.* **38,** 1855–1872 (2011).
41. Samonds, K. E. *et al.* Spatial and temporal arrival patterns of Madagascar's vertebrate fauna explained by distance, ocean currents, and ancestor type. *Proc. Natl Acad. Sci. USA* **109,** 5352–5357 (2012).
42. Samonds, K. E. *et al.* Imperfect isolation: factors and filters shaping Madagascar's extant vertebrate fauna. *PLoS ONE* **8,** e62086 (2013).
43. Evans, S. E., Jones, M. E. H. & Krause, D. W. A giant frog with South American affinities from the Late Cretaceous of Madagascar. *Proc. Natl Acad. Sci. USA* **105,** 2951–2956 (2008).
44. Buckley, G. A., Brochu, C., Krause, D. W. & Pol, D. A pug-nosed crocodyliform from the Late Cretaceous of Madagascar. *Nature* **405,** 941–944 (2000).
45. Forster, C. A., Sampson, S. D., Chiappe, L. M. & Krause, D. W. The theropod ancestry of birds: new evidence from the Late Cretaceous of Madagascar. *Science* **279,** 1915–1919 (1998).
46. Sampson, S. D. *et al.* Predatory dinosaur remains from Madagascar: implications for the Cretaceous biogeography of Gondwana. *Science* **280,** 1048–1051 (1998).
47. Sampson, S. D., Carrano, M. T. & Forster, C. A. A bizarre predatory dinosaur from the Late Cretaceous of Madagascar. *Nature* **409,** 504–506 (2001).
48. Crottini, A. *et al.* Vertebrate time-tree elucidates the biogeographic pattern of major biotic change around the K-T boundary in Madagascar. *Proc. Natl Acad. Sci. USA* **109,** 5358–5363 (2012).
49. Vidal, N. *et al.* Blindsnake evolutionary tree reveals long history on Gondwana. *Biol. Lett.* **6,** 558–561 (2010).
50. Ali, J. R. & Aitchison, J. C. Gondwana to Asia: plate tectonics, paleogeography and the biological connectivity of the Indian sub-continent from the Middle Jurassic through latest Eocene (166–35 Ma). *Earth Sci. Rev.* **88,** 145–166 (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the Université d'Antananarivo, the Madagascar Institut pour la Conservation des Ecosystèmes Tropicaux, and the villagers of the Lac Kinkony Study Area for logistical support of fieldwork; various ministries of the Republic of Madagascar for permission to conduct field research; members of the 2010 field research team for their efforts; J. Thostenson and M. Hill of the American Museum of Natural History Microscopy & Imaging Facility, New York, New York, and J. Diehm and B. Ruether of Avonix Imaging, Plymouth, Minnesota, and various members of the Department of Radiology at Stony Brook University for providing expert assistance in computed tomography scanning; L. Betti-Nash for artwork; J. Neville for photography; D. Pulaski for reconstructing the cranium and building the finite element models; Z.-X. Luo, G. Rougier and A. Weil for their reviews of the paper; and the National Geographic Society (8597-09) and the National Science Foundation (EAR-0446488, EAR-1123642) for funding.

Author Contributions J.R.G. prepared the fossil; J.R.G., S.H., W.L.H. and P.M.O. conducted most of the micro-computed tomography digital preparation; L.J.R. and R.R.R. provided geological data; H.A. provided logistical support; J.R.G., S.H., D.W.K., W.v.K., P.M.O., J.B.R. and J.R.W. provided most of the descriptions and comparisons; E.R.D., A.D.K., D.W.K., E.C.K., W.v.K., P.M.O., J.A.S. and J.R.W. conducted various functional and comparative analyses; S.H., D.W.K., E.R.S. and J.R.W. contributed to the phylogenetic analysis; D.W.K. developed the manuscript, with contributions from all authors.

Author Information *Vintana sertichi* has been assigned the Life Science Identifier (LSID) http://zoobank.org/urn:lsid:zoobank.org:act:B21CC5B2-D550-4D78-BA1F-8319EA663785. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.W.K. (David.Krause@stonybrook.edu).

ARTICLE

doi:10.1038/nature13959

Global diets link environmental sustainability and human health

David Tilman[1,2] & Michael Clark[1]

Diets link environmental and human health. Rising incomes and urbanization are driving a global dietary transition in which traditional diets are replaced by diets higher in refined sugars, refined fats, oils and meats. By 2050 these dietary trends, if unchecked, would be a major contributor to an estimated 80 per cent increase in global agricultural greenhouse gas emissions from food production and to global land clearing. Moreover, these dietary shifts are greatly increasing the incidence of type II diabetes, coronary heart disease and other chronic non-communicable diseases that lower global life expectancies. Alternative diets that offer substantial health benefits could, if widely adopted, reduce global agricultural greenhouse gas emissions, reduce land clearing and resultant species extinctions, and help prevent such diet-related chronic non-communicable diseases. The implementation of dietary solutions to the tightly linked diet–environment–health trilemma is a global challenge, and opportunity, of great environmental and public health importance.

Agriculture is having increasingly strong global impacts on both the environment[1–5] and human health, often driven by dietary changes[6–9]. Global agriculture and food production release more than 25% of all greenhouse gases (GHGs)[2–4], pollute fresh and marine waters with agrochemicals[1,5], and use as cropland or pastureland about half of the ice-free land area of Earth[10]. Despite the intensity and impacts of global agriculture, almost a billion people still suffer from inadequate diets and insecure food supplies[11–13]. Moreover, the global transition towards diets high in processed foods, refined sugars, refined fats, oils and meats has contributed to 2.1 billion people becoming overweight or obese[6,14]. These dietary shifts and resulting increases in body mass indices (BMI) are associated with increased global incidences of chronic non-communicable diseases, especially type II diabetes, coronary heart disease and some cancers[7–9,15–22], which are predicted to become two-thirds of the global burden of disease if dietary trends continue[9,16,17]. In China, for instance, as incomes increased and diets changed[20], the incidence of type II diabetes increased from <1% of its population in 1980 to 10% in 2008, partly because type II diabetes occurs at lower BMI levels and earlier in an individual's life in Asian than in western populations[9]. Moreover, diet-driven increases in global food demand[7,8,12,23] and increases in population are leading to clearing of tropical forests, savannas and grasslands[1,5,23], which threatens species with extinction[1,3–5,23–25].

Because it directly links and negatively affects human and environmental health, the global dietary transition is one of the great challenges facing humanity. Meaningful solutions will not be easily achieved. Solutions will require analyses of the quantitative linkages between diets, the environment and human health, on which we focus here, and the efforts of nutritionists, agriculturists, public health professionals, educators, policy makers and food industries.

Here we compile and analyse global-level data to quantify relationships among diet, environmental sustainability and human health, evaluate potential future environmental impacts of the global dietary transition and explore some possible solutions to the diet–environment–health trilemma (Methods and Supplementary Information). To do so, we first expand on earlier food lifecycle analyses[24,25] (LCAs) by searching for all published LCAs of GHG emissions of food crop, livestock, fishery and aquaculture production systems that delimited the full 'cradle to farm gate' portion of the food/crop lifecycle. Next we use about 50 years of data for 100 of the world's more populous nations to analyse global dietary trends and their drivers, then use this information to forecast future diets should past trends continue. To quantify effects of alternative diets on mortality and on type II diabetes, cancer and chronic coronary heart disease, we compile and summarize results of studies encompassing ten million person-years of observations on diet and health. Finally, we combine these relationships with projected increases in global population to forecast global environmental implications of current dietary trajectories and to calculate the environmental benefits of diets associated with lower incidences of chronic non-communicable diseases.

Lifecycle environmental impacts of foods

Dietary composition strongly influences GHG emissions[2,24–27]. The 120 LCA publications that met our criteria report a total of 555 LCA analyses on 82 types of crops and animal products, allowing us to calculate diet-related GHG emissions per gram protein, per kilocalorie and per serving from 'cradle to farm gate' (Fig. 1; Methods, Extended Data Tables 1–3). We express emissions as CO_2 warming equivalents, in grams (g) or gigatonnes (Gt) of CO_2 carbon equivalents (CO_2-C_{eq}).

GHG emissions vary widely among foods (Fig. 1; Extended Data Table 3 lists means, s.e.m. and number of data points). As is well known, relative to animal-based foods, plant-based foods have lower GHG emissions. This difference can be large; the largest we found was that ruminant meats (beef and lamb) have emissions per gram of protein that are about 250 times those of legumes (Extended Data Table 3; Student's *t*-test comparison of means: $P < 0.0001$). Eggs, dairy, non-trawling seafood, traditional (non-recirculating) aquaculture, poultry and pork all have much lower emissions per gram of protein than ruminant meats (Tukey range test comparing ruminant meats with each other item: $P < 0.0001$ for each comparison). However, when sustainably grazed on lands unsuitable for cropping and fed crop residues, ruminant dairy and meat production can increase food security, dietary quality, and provide environmental benefits via nutrient cycling[28,29]. How a given food is produced can also affect emissions. Seafood caught by trawling, in which nets are often dragged across the ocean floor, has emissions per gram of protein about 3 times those of non-trawling seafood (Fig. 1; Extended Data Table 3; *t*-test mean comparison: $P = 0.017$). Items within the same food group can

[1]Department of Ecology, Evolution and Behavior, University of Minnesota, St Paul, Minnesota 55108, USA. [2]Bren School of Environmental Science and Management, University of California Santa Barbara, California 93106, USA.

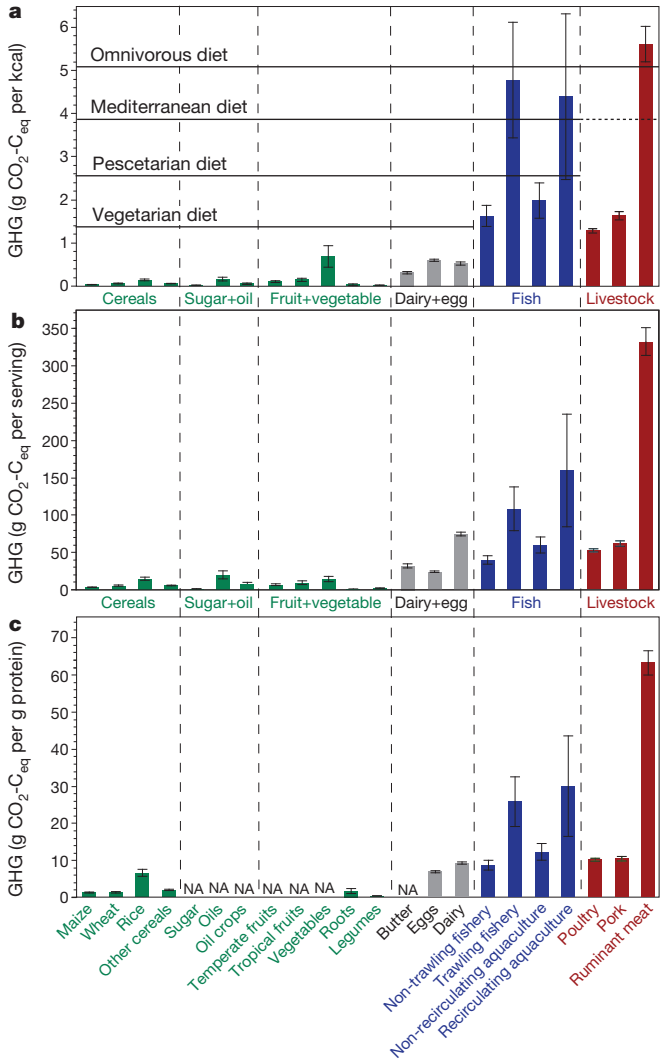


Figure 1 | Lifecycle GHG emissions (CO_2-C_{eq}) for 22 different food types. The data are based on an analysis of 555 food production systems: **a**, per kilocalorie; **b**, per United States Department of Agriculture (USDA)-defined serving; **c**, per gram of protein. The mean and s.e.m. are shown for each case. Extended Data Tables 1–3 list data sources, items included in each of the 22 food types and show the mean, s.e.m. and number of data points for each bar, respectively. NA, not applicable.

also differ. For instance, among cereal grains, wheat has a fifth the GHG emissions per g protein of rice (*t*-test comparison: $P = 0.002$).

Finally, to understand its environmental impacts, it is important to know the nutritional needs that a food meets and how much is consumed to do so. Fruits and vegetables are important sources of micronutrients, antioxidants and fibre. Unlike root crops and legumes, which are calorie-dense or protein-dense, most vegetables are not primarily consumed for calories or protein and should be evaluated by emissions per serving. For instance, 20 servings of vegetables have less GHG emissions than one serving of beef (Fig. 1b). However, fish and meats, which are high in protein, are also nutritionally dense foods that provide essential fatty acids, minerals and vitamins[28,29], and can have relatively low GHG emissions if eaten in moderation. Finally, the nutritional value of some foods can depend on how they are produced. For instance, in comparison to grain-fed cattle, grass-fed beef and dairy have nutritionally superior fatty acid and vitamin content[30].

Global dietary change

Although diets differ within and among nations and regions for a variety of climatic, cultural and historic reasons, diets have been changing in fairly consistent ways as incomes and urbanization have increased globally during the past five decades[6–9]. This dietary transition has many components, but, in broad outline, its magnitude and global nature are illustrated by trends in per capita demand for meat, empty calories and total calories (Fig. 2), where demand is defined as food brought into a household.

As annual incomes (per capita real gross domestic product, GDP) increased from 1961 to 2009, there were concomitant increases in per capita daily demand for meat protein (Fig. 2a) within and among eight economically based groups of nations[23] (Extended Data Table 4). In 2009, the richest 15 nations (Group A; Fig. 2a) had a 750% greater per capita demand for meat protein from ruminants, seafood, poultry and pork than the 24 poorest nations (Group F). Total protein demand also increased with income, but legume protein demand decreased as animal protein demand increased. India, a nation with low rates of meat consumption, is the major exception to an otherwise global trend in the income-dependence of demand for meat protein (Fig. 2a). China initially had meat demand increase more rapidly with income than Groups A–F, but was similar to them by 2009.

A second trend within and among economic groups is the income-dependent increase in demand for 'empty calories', here defined as calories from refined fats, refined sugars, alcohols and oils (Fig. 2b). In 2009, Group A nations had an average per capita empty calorie demand of 1,400 kcal per day, whereas demand was 285 kcal per day for Group F. The exception, China, is on an increasing but lower trajectory (Fig. 2b).

A third trend is that total per capita caloric demand also increased with income (Fig. 2c), with China falling below the fitted trend, and Group A being above it. Because some food brought into homes (demand) is wasted[13], and the proportion wasted tends to increase with per capita GDP[31], actual per capita consumption of meat, empty calories and total calories may be about 20%–25% lower than demand for the Group A nations and about 5% lower in Group F nations. This suggests that, in nations with per capita GDP above approximately $12,000 per year (in 1990$), per capita total caloric consumption may be about 500 kcal per day greater than needed nutritionally.

In total, annual data for 1961 to 2009 for China, India and six income-based groups of nations show that global dietary changes are associated with increased income (Fig. 2), which is itself associated with urbanization and industrial food production[20]. When these trends are combined with forecasts of per capita income for the coming decades, we estimate that, relative to the average global diet of 2009, the 2050 global-average per capita income-dependent diet would have 15% more total calories and 11% more total protein, with dietary composition shifting to having 61% more empty calories, 18% fewer servings of fruits and vegetables, 2.7% less plant protein, 23% more pork and poultry, 31% more ruminant meat, 58% more dairy and egg and 82% more fish and seafood.

Diet and human health

Diet is an important determinant of human health. Many of the world's poorest people have inadequate diets, and would have improved health were their diets to include more essential fatty acids, minerals, vitamins and protein from fish and meats and added calories and protein from other nutritionally appropriate sources[12,29]. In contrast, diets of many people with moderate and higher incomes are shifting in ways (Fig. 2) associated with increases in non-communicable diseases[6,7] including type II diabetes[9,19], coronary heart disease[21] and cancer[21], and with higher all-cause mortality rates[18,22].

A point of contrast to the detrimental health impacts of emerging global diets is provided by the benefits reported for three well-studied alternative diets. Here we summarize results from ten million person-years of observations across eight study cohorts[32–39] (Methods; Extended Data Table 5). For each cohort we use reported health outcome effect sizes that had been calculated after statistical control for potentially confounding variables to compare disease incidence rates of individuals who consumed typical omnivorous diets with those who had diets classified as Mediterranean, pescetarian or vegetarian (Fig. 1a). These diets

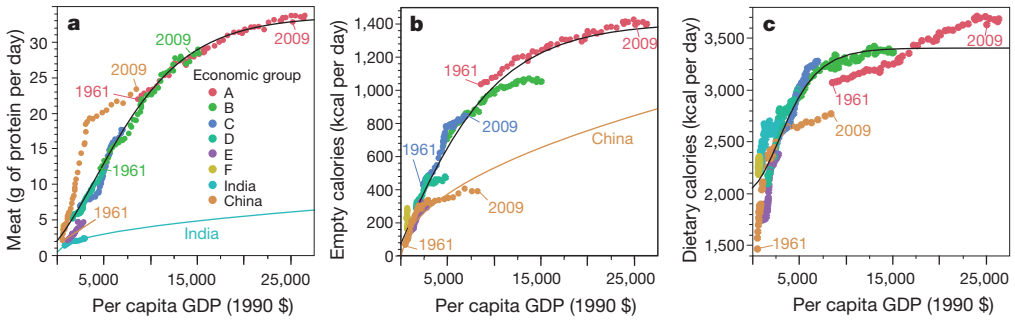


Figure 2 | Dietary trends and income. Dependence of per capita daily dietary demand for: **a**, meat protein; **b**, refined sugars+refined animal fats+oils+alcohol; and **c**, calories on per capita gross domestic product (GDP measured in 1990 International Dollars). Each point is an annual datum for 1961 to 2009 for India, China, and six economic groups containing 98 other nations (Extended Data Table 4). Fitted curves were used to forecast 2050 income-dependent demand.

have different compositions. A vegetarian diet consists of grains, vegetables, fruits, sugars, oils, eggs and dairy, and generally not more than one serving per month of meat or seafood. A pescetarian diet is a vegetarian diet that includes seafood. A Mediterranean diet is rich in vegetables, fruit and seafood and includes grains, sugars, oils, eggs, dairy and moderate amounts of poultry, pork, lamb and beef. Omnivorous diets, such as the 2009 global-average diet and the income-dependent 2050 diet, include all food groups.

Relative to conventional omnivorous diets, across the three alternative diets incidence rates of type II diabetes were reduced by 16%–41% and of cancer by 7%–13%, while relative mortality rates from coronary heart disease were 20%–26% lower and overall mortality rates for all causes combined were 0%–18% lower (Fig. 3). This summary illustrates the magnitudes of the health benefits associated with some widely adopted alternative diets. The alternative diets tend to have higher consumption of fruits, vegetables, nuts and pulses and lower empty calorie and meat consumption than the 2009 average global diet and the 2050 income-dependent diet (Extended Data Fig. 1). Our analyses are not designed to compare the health impacts of the three alternative diets with each other, nor to imply that other diets might not provide health benefits superior to these three diets. Indeed, the reported impacts of individual foods, such as deleterious impacts from sugars[40] and processed meats[19,22], and benefits from nuts and olive oil[41], suggest that variants of these three diets may offer added health benefits, as may other diets.

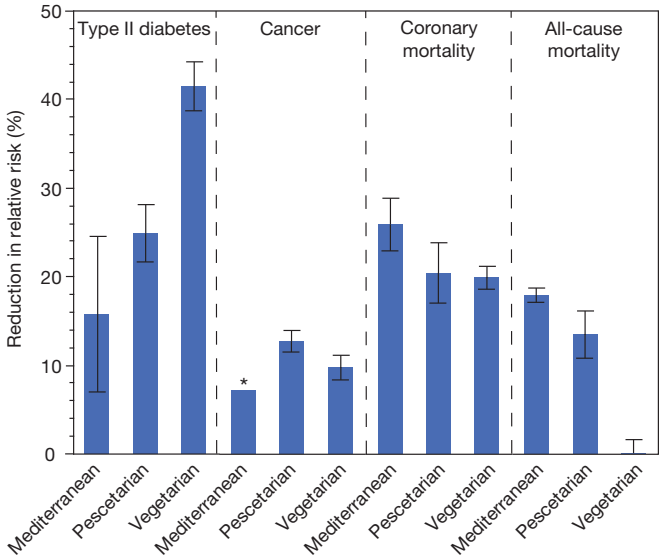


Figure 3 | Diet and health. Diet-dependent percentage reductions in relative risk of type II diabetes, cancer, coronary heart disease mortality and of all-cause mortality when comparing each alternative diet (Mediterranean, pescetarian and vegetarian) to its region's conventional omnivorous diet (Methods). Results are based on cohort studies[32–39]. The mean and s.e.m. values shown are weighted by person-years of data for each study. Number of studies for each bar are, from left to right, 3, 2, 2, 1, 2, 2, 4, 2, 5, 13, 2 and 4. *Cancer in Mediterranean diets is from a single study so no s.e.m. is shown.

Environmental impacts of diets

GHG emissions are highly dependent on diet[24–27,42–44]. Even foods that provide similar nutrition and have similar impacts on health can have markedly different lifecycle environmental impacts. Using LCA emission data, we calculated annual per capita GHG emissions from food production ('cradle to farm gate') for the 2009 global-average diet, for the global-average income-dependent diet projected for 2050, and for Mediterranean, pescetarian and vegetarian diets (Fig. 4a). Global-average per capita dietary GHG emissions from crop and livestock production would increase 32% from 2009 to 2050 if global diets changed in the income-dependent ways illustrated in Fig. 2. All three alternative diets could reduce emissions from food production below those of the projected 2050 income-dependent diet (Fig. 4a), with per capita reductions being 30%, 45% and 55% for the Mediterranean, pescetarian and vegetarian diets, respectively. However, minimizing environmental impacts does not necessarily maximize human health. Prepared items high in sugars, fats or carbohydrates can have low GHG emissions (Fig. 1) but be less healthy than foods they displace[20]. Solutions to the diet–environment–health trilemma should seek healthier diets that have low GHG emissions rather than diets that might minimize GHG emissions.

Changes towards healthier diets can have globally significant GHG benefits (Fig. 4b). From 2009 to 2050 global population is projected to increase by 36% (ref. 10). When combined with the projected 32% increase in per capita emissions from income-dependent global dietary shifts, the net effect is an estimated 80% increase in global GHG emissions from food production (from 2.27 to 4.1 Gt yr^{-1} of CO_2-C_{eq}).This increase of 1.8 Gt yr^{-1} is equivalent to total 2010 global transportation emissions[3]. In contrast, there would be no net increase in food production emissions if by 2050 the global diet had become the average of the Mediterranean, pescetarian and vegetarian diets (Fig. 4b).

Future global land clearing for agriculture could threaten species with extinction[1,5] and release GHG beyond that from food production. However, the extent of such land clearing is uncertain, variously projected to total from 0 to 10^9 hectares[5,23,45,46] by 2050, perhaps because of uncertainties about the future values of five factors: crop yields, agricultural and food waste, livestock yields from pastures, animal feed use efficiency and agricultural trade. Here we focus not on forecasting the absolute amount of cropland needed in 2050, but on estimating across many scenarios (243 combinations of three values for each of the five factors; Methods) the differential impacts of diets on global cropland. The alternative scenarios forecast a range of changes in cropland from 2009 to 2050 for each diet (Fig. 4c). For each scenario we calculated the difference between projected 2050 land demands of the income-dependent diet

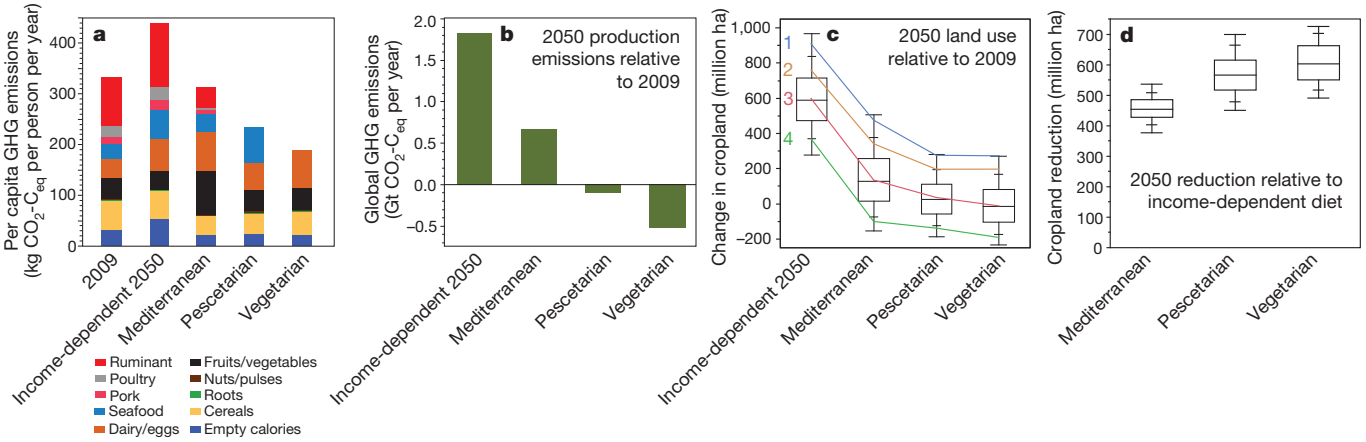


Figure 4 | Effect of diets on GHG emissions and cropland. **a**, Per capita food production GHG emissions for five diets (2009 global-average, 2050 global income-dependent, Mediterranean, pescetarian and vegetarian). **b**, **c**, Forecasted 2009 to 2050 changes (2009 value set to 0) in global food emissions (**b**), and cropland for each diet (Methods; alternative scenarios, such as lines 1-4, have fairly parallel trends) (**c**). **d**, 2050 global cropland reductions from alternative diets relative to income-dependent diet. The box and whisker plots (**c**, **d**) show mean (centre line) and percentiles below (2.5th, 10th, 25th) and above it (75th, 90th, and 97.5th) based on 243 scenarios.

and of each alternative diet (Fig. 4d). Across these scenarios, the income-dependent diet requires from 370 to 740 million hectares more cropland than the alternative diets, and averages 540 million hectares more (Fig. 4d). These results suggest that shifts towards healthier diets could substantially decrease future agricultural land demand and clearing, as could improvements in the five factors (Extended Data Table 6). Land clearing also leads to GHG emissions. Clearing 540 million hectares from 2010 to 2050 would release about 0.6 Gt yr^{-1} of CO_2-C_{eq}.

In addition to dietary shifts, other changes will be needed for agriculture to become environmentally sustainable[13,23,28–31,47–49]. If, by 2050, all forms of crop and food wastage[13,31] were globally reduced by 50%, food production emissions could be reduced by about 0.5 Gt yr^{-1} of CO_2-C_{eq} relative to the 2050 income-dependent diet. Increases in use efficiencies of animal feeds (from those of Extended Data Table 7), fertilizer and irrigation, and improvements in pasture management and aquaculture would increase food production, decrease GHG emissions and improve water quality[28,29,47–49]. Increases in yields of under-yielding nations could also reduce emissions[23]. Climate change, though, can affect yields[50], which could in turn have an impact on agricultural GHG emissions and land clearing.

Discussion

Dietary choices link environmental sustainability and human health. Current dietary trajectories (Fig. 2) are greatly increasing global incidences of type II diabetes, cancer and coronary heart disease. These dietary changes are causing globally significant increases in GHG emissions and contributing to land clearing. Although this pattern does not mean that healthier diets are necessarily more environmentally beneficial, nor that more environmentally beneficial diets are necessarily healthier, there are many alternative dietary options that should substantially improve both human and environmental health.

Our analyses demonstrate that there are plausible solutions to the diet–environment–health trilemma, diets already chosen by many people that, if widely adopted, would offer global environmental and public health benefits. Clearly, to appeal to specific segments of the global population, other such diets should also be developed. The health benefits of adopting such diets could be substantial. Chronic diet-related non-communicable diseases are affecting an increasing number of children and adults in all but the poorest nations. Nations ranging from China and India to Mexico, Nigeria and Tunisia are in the midst of this increasing disease incidence[17]. Unless the nutrition transition that is under way is changed, diabetes, chronic heart disease and other diet-related chronic non-communicable diseases will become the dominant global disease burden, often affecting even the poorer members of poorer nations for whom appropriate health care is unavailable[16,17].

The dietary choices that individuals make are influenced by culture, nutritional knowledge, price, availability, taste and convenience, all of which must be considered if the dietary transition that is taking place is to be counteracted. The evaluation and implementation of dietary solutions to the tightly linked diet–environment–health trilemma is a global challenge, and opportunity, of great environmental and public health importance.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 20 April; accepted 13 October 2014.
Published online 12 November 2014.

1. Matson, P. A., Parton, W. J., Power, A. G. & Swift, M. J. Agricultural intensification and ecosystem properties. *Science* **277,** 504–509 (1997).
2. Steinfeld, H. *et al. Livestock's Long Shadow* (FAO, 2006).
3. Edenhofer, O. *et al. Climate Change 2014: Mitigation of Climate Change Technical Summary* (Intergovernmental Panel on Climate Change, 2014).
4. Tubiello, F. N. *et al. Agriculture, Forestry and other Land Use Emissions by Sources and Removals by Sinks* (FAO Statistics Division, ESS/14-02, 2014).
5. Tilman, D. *et al.* Forecasting agriculturally driven global environmental change. *Science* **292,** 281–284 (2001).
6. Popkin, B. M., Adair, L. S. & Ng, S. W. Global nutrition transition and the pandemic of obesity in developing countries. *Nutr. Rev.* **70,** 3–21 (2012).
7. Popkin, B. M. The nutrition transition in low-income countries: an emerging crisis. *Nutr. Rev.* **52,** 285–298 (1994).
8. Drewnowski, A. & Popkin, B. M. The nutrition transition: new trends in the global diet. *Nutr. Rev.* **55,** 31–43 (1997).
9. Hu, F. B. Globalization of diabetes: the role of diet, lifestyle, and genes. *Diabetes Care* **34,** 1249–1257 (2011).
10. Food and Agriculture Organization of the United Nations. http://faostat.fao.org (FAO, 2013).
11. Smil, V. *Feeding the World: a Challenge for the Twenty-First Century* (MIT Press, 2000).
12. FAO. Global agriculture towards 2050. In *How to Feed the World 2050* 1–10 (FAO, 2009).
13. Godfray, H. C. J. *et al.* Food security: the challenge of feeding 9 billion people. *Science* **327,** 812–818 (2010).
14. Ng, M. *et al.* Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* **6736,** 1–16 (2014).
15. Willett, W. C. *et al.* Mediterranean diet pyramid: a cultural model for healthy eating. *Am. J. Clin. Nutr.* **61,** 1402S–1406S (1995).
16. Chopra, M., Galbraith, S. & Darnton-Hill, I. A global response to a global problem: the epidemic of overnutrition. *Bull. World Health Organ.* **80,** 952–958 (2002).
17. Nishida, C. *et al.* Diet, nutrition and the prevention of chronic diseases: report of a joint WHO/FAO expert consultation. *Public Health Nutr.* **7,** 245–250 (2004).
18. Singh, P. N., Sabaté, J. & Fraser, G. E. Does low meat consumption increase life expectancy in humans? *Am. J. Clin. Nutr.* **78,** 526S–532S (2003).

19. Aune, D., Ursin, G. & Veierød, M. B. Meat consumption and the risk of type 2 diabetes: a systematic review and meta-analysis of cohort studies. *Diabetologia* **52,** 2277–2287 (2009).
20. Kearney, J. Food consumption trends and drivers. *Phil. Trans. R. Soc. B* **365,** 2793–2807 (2010).
21. Huang, T. *et al.* Coronary heart disease mortality and cancer incidence in vegetarians: a meta-analysis and systematic review. *Ann. Nutr. Metab.* **60,** 233–240 (2012).
22. Pan, A. *et al.* Red meat consumption and mortality: results from 2 prospective cohort studies. *Arch. Intern. Med.* **172,** 555–563 (2012).
23. Tilman, D., Balzer, C., Hill, J. & Befort, B. L. Global food demand and the sustainable intensification of agriculture. *Proc. Natl Acad. Sci. USA* **108,** 20260–20264 (2011).
24. de Vries, M. & de Boer, I. J. M. Comparing environmental impacts for livestock products: a review of life cycle assessments. *Livest. Sci.* **128,** 1–11 (2010).
25. Nijdam, D., Rood, T. & Westhoek, H. The price of protein: review of land use and carbon footprints from life cycle assessments of animal food products and their substitutes. *Food Policy* **37,** 760–770 (2012).
26. Eshel, G. & Martin, P. A. Diet, energy, and global warming. *Earth Interact.* **10,** 1–17 (2006).
27. Marlow, H. J. *et al.* Diet and the environment: does what you eat matter? *Am. J. Clin. Nutr.* **89,** 1699–1703 (2009).
28. Eisler, M. C. *et al.* Steps to sustainable livestock. *Nature* **507,** 32–34 (2014).
29. Smith, J. *et al.* Beyond milk, meat and eggs: role of livestock in food and nutrition security. *Anim. Front.* **3,** 6–13 (2013).
30. Daley, C. A., Abbott, A., Doyle, P. S., Nader, G. & Larson, S. A review of fatty acid profiles and antioxidant content in grass-fed and grain-fed beef. *Nutr. J.* **9,** 10 (2010).
31. Gustavsson, J., Cederberg, C., van Otterdiijk, R. & Meybeck, A. *Global Food Losses and Food Waste* (FAO, 2011).
32. Snowdon, D. A., Phillips, R. L. & Fraser, G. E. Meat consumption and fatal ischemic heart disease. *Prev. Med.* **13,** 490–500 (1984).
33. Key, T. J., Thorogood, M., Appleby, P. N. & Burr, M. L. Dietary habits and mortality in 11,000 vegetarians and health conscious people: results of a 17 year follow up. *Br. Med. J.* **313,** 775–779 (1996).
34. Mann, J. I., Appleby, P. N., Key, T. J. & Thorogood, M. Dietary determinants of ischaemic heart disease in health conscious individuals. *Heart* **78,** 450–455 (1997).
35. Lagiou, P. *et al.* Mediterranean dietary pattern and mortality among young women: a cohort study in Sweden. *Br. J. Nutr.* **96,** 384–392 (2006).
36. Mitrou, P. N. *et al.* Mediterranean dietary pattern and prediction of all-cause mortality in a US population. *Arch. Intern. Med.* **167,** 2461–2468 (2007).
37. Brunner, E. J. *et al.* Dietary patterns and 15-y risks of major coronary events, diabetes, and mortality. *Am. J. Clin. Nutr.* **87,** 1414–1421 (2008).
38. Martínez-González, M. A. *et al.* Adherence to Mediterranean diet and risk of developing diabetes: prospective cohort study. *Br. Med. J.* **336,** 1348–1351 (2008).
39. Fung, T. T. *et al.* Mediterranean diet and incidence of and mortality from coronary heart disease and stroke in women. *Circulation* **119,** 1093–1100 (2009).
40. Yang, Q. *et al.* Added sugar intake and cardiovascular diseases mortality among US adults. *JAMA Intern. Med.* **174,** 516–524 (2014).
41. Buckland, G. *et al.* Olive oil intake and mortality within the Spanish population (EPIC-Spain). *Am. J. Clin. Nutr.* **96,** 142–149 (2012).
42. Stehfest, E. *et al.* Climate benefits of changing diet. *Clim. Change* **95,** 83–102 (2009).
43. Popp, A., Lotze-Campen, H. & Bodirsky, B. Food consumption, diet shifts and associated non-CO_2 greenhouse gases from agricultural production. *Glob. Environ. Change* **20,** 451–462 (2010).
44. Westhoek, H. *et al.* Food choices, health and environment: effects of cutting Europe's meat and dairy intake. *Glob. Environ. Change* **26,** 196–205 (2014).
45. Alexandratos, N. & Bruinsma, J. *World Agriculture Towards 2030/2050: The 2012 Revision* Ch. 4 (ESA/12-03, FAO, 2012).
46. Schmitz, C. *et al.* Land-use change trajectories up to 2050: insights from a global agro-economic model comparison. *Agric. Econ.* **45,** 69–84 (2014).
47. Herrero, M. *et al.* Biomass use, production, feed efficiencies, and greenhouse gas emissions from global livestock systems. *Proc. Natl Acad. Sci. USA* **110,** 20888–20893 (2013).
48. Havlík, P. *et al.* Climate change mitigation through livestock system transitions. *Proc. Natl Acad. Sci. USA* **111,** 3709–3714 (2014).
49. Chen, X.-P. *et al.* Integrated soil-crop system management for food security. *Proc. Natl Acad. Sci. USA* **108,** 6399–6404 (2011).
50. Hatfield, J. L. *et al.* Climate impacts on agriculture: implications for crop production. *Agron. J.* **103,** 351–370 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank M. Burgess, A. Clark and E. Hallström for their comments, K. Thompson for assistance with data collection, editing, and creating figures, and the LTER programme of the US National Science Foundation and the University of Minnesota Foundation for support.

Author Contributions D.T. conceived this project and M.C. assembled data; both M.C. and D.T. analysed data and wrote the paper.

Author Information All data used in our analyses are publicly available from the original sources that we list, and are provided in the Supplementary Information. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.T. (tilman@umn.edu).

ARTICLE

doi:10.1038/nature13953

A positional Toll receptor code directs convergent extension in *Drosophila*

Adam C. Paré[1], Athea Vichas[1], Christopher T. Fincher[1], Zachary Mirman[1], Dene L. Farrell[1], Avantika Mainieri[1] & Jennifer A. Zallen[1]

Elongation of the head–to–tail body axis by convergent extension is a conserved developmental process throughout metazoans. In *Drosophila*, patterns of transcription factor expression provide spatial cues that induce systematically oriented cell movements and promote tissue elongation. However, the mechanisms by which patterned transcriptional inputs control cell polarity and behaviour have long been elusive. We demonstrate that three Toll family receptors, Toll–2, Toll–6 and Toll–8, are expressed in overlapping transverse stripes along the anterior–posterior axis and act in combination to direct planar polarity and polarized cell rearrangements during convergent extension. Simultaneous disruption of all three receptors strongly reduces actomyosin–driven junctional remodelling and axis elongation, and an ectopic stripe of Toll receptor expression is sufficient to induce planar polarized actomyosin contractility. These results demonstrate that tissue–level patterns of Toll receptor expression provide spatial signals that link positional information from the anterior–posterior patterning system to the essential cell behaviours that drive convergent extension.

A central question in developmental biology is how the diverse structures of multicellular tissues are generated on a cellular and molecular level. Convergent extension, in which a tissue narrows along one axis and lengthens in a perpendicular direction, is a conserved tissue remodelling process that elongates the head-to-tail body axis of many animals. Cell intercalation provides the driving force for convergent extension in frogs, fish, flies, chicks and mice[1–5]. This process is characterized by a striking directionality in which hundreds of cells align their movements along a common axis. In epithelial tissues, cell intercalation is mediated by spatially regulated actomyosin contractility, which induces locally oriented cell rearrangements that produce concerted elongation at the tissue scale[6–10]. This mechanism was first discovered in *Drosophila*[6–10], and has since been shown to promote convergent extension in the vertebrate neural plate, primitive streak, kidney and notochord[11–16]. In *Drosophila* and *Xenopus*, the spatial cues that align cell movements with the tissue axes are not cell-intrinsic or long-range secreted signals. Instead, contact-dependent signals provide the critical spatial inputs that orient cell intercalation[6,17,18]. In the *Drosophila* embryo, these inputs are mediated by the pair-rule transcription factors Eve and Runt, components of the embryonic anterior–posterior (AP) patterning system that are expressed in transverse stripes along the AP axis[19]. When these striped patterns are disrupted, either in *eve* or *runt* mutants or in embryos overexpressing Eve or Runt at high levels, intercalary behaviours are reduced and misoriented[8,17,20] and the actomyosin contractile machinery becomes mislocalized within cells[6,8,21]. Ectopic Eve or Runt expression perpendicular to their normal stripes is sufficient to reorient planar polarity in intercalating cells[6], demonstrating that spatial patterns of Eve and Runt activity provide instructive polarity cues. However, the connection between transcriptional information provided by striped patterns of Eve and Runt activity and the effector molecules that generate polarized cell behaviour during convergent extension has long been elusive[2,22].

A positional code of Toll receptor stripes

To identify the targets of Eve and Runt that direct cell behaviour during convergent extension, we performed RNA sequencing on *Drosophila* embryos co-injected with *eve* and *runt* double-stranded RNAs (dsRNAs). Compared with water-injected controls, 42 genes were differentially expressed ($P < 0.01$, Extended Data Fig. 1a–c and Supplementary Table 1). As Eve and Runt can function as transcriptional repressors, we focused on the 24 genes that were significantly upregulated by *eve/runt* RNA interference (RNAi). This group included *Toll-8* (also known as *Tollo*), which encodes a single-pass transmembrane protein containing 27 extracellular leucine-rich repeats (LRRs) and a conserved cytoplasmic Toll/interleukin-1 receptor (TIR) domain. The related gene *Toll-2* (also known as *18-wheeler*) was also upregulated by *eve/runt* RNAi. Toll-2 and Toll-8 belong to the Toll receptor family that regulates innate immunity in arthropods and vertebrates[23–27]. The founding member of this family, Toll, is essential for dorsal-ventral patterning in *Drosophila*[28], and the paralogues *Toll-2*, *Toll-6*, *Toll-7* and *Toll-8* are expressed in stripes in patterns reminiscent of the pair-rule genes[29–31]. As members of this family can influence cell adhesion and epithelial morphology[31–35], we focused on Toll receptors as candidate effectors of Eve and Runt that could regulate polarized cell behaviour during convergent extension.

To determine whether the striped expression of Toll family receptors requires Eve and Runt activity, we analysed their expression by fluorescence *in situ* hybridization[36]. In wild-type embryos, *Toll-2* is expressed in 13 stripes in the germband, of which the odd-numbered stripes coincide with Runt (Fig. 1a, d and Extended Data Fig. 2a, d)[29–31]. *Toll-6* and *Toll-8* are each expressed in 6 partially overlapping stripes, with *Toll-6* expressed anterior to the stripes of high *Toll-2* expression and *Toll-8* expressed between them (Fig. 1b, c, e, f, m, n and Extended Data Fig. 2b, c, e, f). *Toll-7* was detected at low levels during axis elongation (Extended Data Fig. 1d and Supplementary Table 1). The 13 *Toll-2* stripes collapsed into 6 broad stripes in *eve* and *runt* mutants (Fig. 1g, j). *Toll-6* levels were reduced and *Toll-8* was nearly absent in *eve* mutants (Fig. 1h, i), whereas *Toll-6* and *Toll-8* were upregulated and more uniformly expressed in *runt* mutants (Fig. 1k, l). The Toll-8 protein fused to YFP and expressed from its endogenous regulatory sequences localized to the plasma membrane, with cells in the middle of each stripe showing stronger signal with no obvious overall planar polarity (Fig. 1o). Therefore, neighbouring cells along the AP axis express different combinations of *Toll-2*, *Toll-6* and *Toll-8* in an Eve- and Runt-dependent pattern (Fig. 1p).

[1]Howard Hughes Medical Institute and Developmental Biology Program, Sloan Kettering Institute, New York, New York 10065, USA.

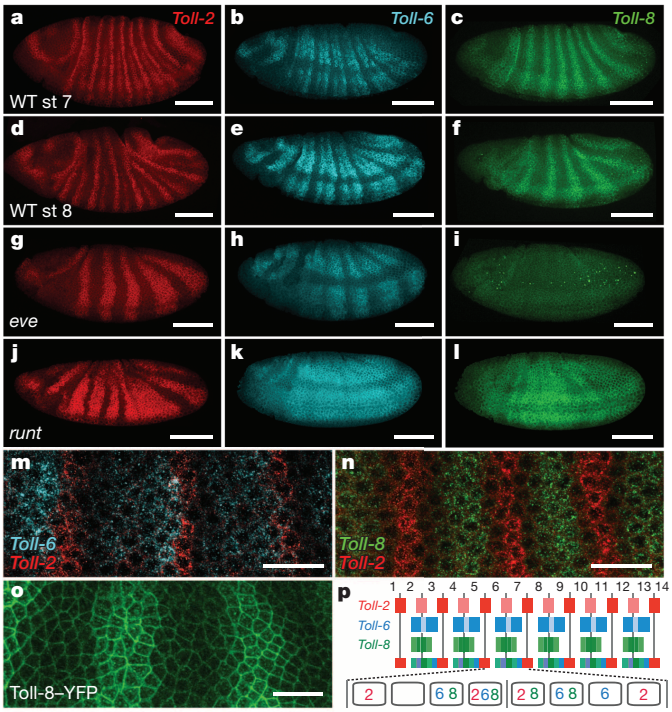


Figure 1 | Cells express different combinations of *Toll-2*, *Toll-6* and *Toll-8* along the anterior–posterior axis. **a**–**n**, *Toll-2* (red), *Toll-6* (cyan) and *Toll-8* (green) mRNA expression in wild-type (WT) (**a**–**f**, **m**, **n**), *eve* mutant (**g**–**i**), and *runt* mutant (**j**–**l**) embryos during early (stage 7, **a**–**c**, **m**, **n**) and mid-elongation (stage 8, **d**–**l**). **m**, **n**, *Toll-6* (cyan) is expressed anterior to the strong *Toll-2* stripes (red) and *Toll-8* (green) is expressed between them. **o**, Toll-8–YFP protein in a stage 7 *Toll-8* mutant. **p**, Schematic of *Toll-2*, *Toll-6* and *Toll-8* expression. Numbers, parasegments; vertical lines, parasegmental boundaries. Anterior left, ventral down. Scale bars, 100 μm (**a**–**l**), 20 μm (**m**–**o**).

Toll receptors direct cell intercalation

We next investigated whether Toll-2, Toll-6 and Toll-8 are required for convergent extension, as predicted for the targets of Eve and Runt that control cell behaviour. The wild-type germband epithelium doubles in length along the AP axis within the first 30 min of elongation (2.00 ± 0.07-fold increase in length) (Fig. 2a–c). Axis elongation occurred normally in *Toll-8* single mutants (Fig. 2c and Extended Data Fig. 3a, b, f). Therefore, we postulated that multiple Toll receptors act together to regulate cell behaviour during elongation. To disrupt multiple Toll receptors simultaneously, we injected dsRNAs that specifically target *Toll-2* and *Toll-6* into *Toll-8* null mutant embryos (Extended Data Fig. 1e, f). Embryos defective for any one receptor elongated to a wild-type extent (Fig. 2b, c and Extended Data Fig. 3a, b, f). By contrast, axis elongation was reduced by nearly 20% in embryos defective for *Toll-2* and *Toll-6* (1.83 ± 0.03-fold, $P < 0.02$) and nearly 40% in embryos defective for *Toll-2*, *Toll-6* and *Toll-8* (1.61 ± 0.04-fold, $P < 0.001$) (Supplementary Video 1), similar to *eve* and *runt* mutants (1.68 ± 0.05-fold in *eve* and 1.64 ± 0.02-fold in *runt*, $P < 0.01$) (Fig. 2c and Extended Data Fig. 3d, f). In addition, we used TAL effector nucleases (TALENs)[37] to generate embryos that completely lack *Toll-2*, *Toll-6* and *Toll-8*, and found that *Toll-2,6,8* triple mutants display a significant reduction in axis elongation (Fig. 2b, c, Extended Data Figs 3e, f and 4 and Supplementary Video 2). These results demonstrate that Toll-2, Toll-6 and Toll-8 act in combination to regulate axis elongation.

Cell intercalation is the primary mechanism driving axis elongation in *Drosophila*[7,8,17]. To determine whether Toll receptors are required for cell intercalation, we used automated methods to track cell behaviour in time-lapse movies[21,38]. In embryos defective for any one Toll family receptor, the frequency of cell intercalation was similar to wild type (Fig. 2f and Extended Data Fig. 3a, b, g–i). By contrast, cell intercalation was reduced by 17% in embryos defective for *Toll-2* and *Toll-6* ($P < 0.03$), 19% in embryos defective for *Toll-6* and *Toll-8* ($P < 0.02$), and more than 30% in *Toll-2,6,8* triple mutants ($P < 0.001$), accompanied by slower edge contraction (Fig. 2e, f and Extended Data Fig. 3c–e, g–j). *Toll-2,6,8* triple mutants were similar to *runt* mutants, although not quite as severe as *eve* mutants (Fig. 2e, f; Extended Data Fig. 3e, g–i). These results demonstrate that Toll-2, Toll-6 and Toll-8 promote cell intercalation during axis elongation.

For cell rearrangements to produce tissue elongation, intercalation must occur directionally through the contraction of interfaces between anterior and posterior neighbours (AP edges) and the formation of interfaces between dorsal and ventral neighbours (DV edges) (Fig. 2d)[7,8]. Contracting edges were correctly oriented in all Toll receptor-defective embryos (Extended Data Fig. 3k). By contrast, in more than one-third of cell rearrangements in *Toll-2,6,8* mutants, new edges failed to form, were unstable, or formed in the wrong direction (36 ± 4% of edges in *Toll-2,6,8* vs 9.5 ± 0.3% in wild type, $P < 0.0001$), similar to the defects in *eve* and *runt* mutants (34 ± 4% in *eve* and 37 ± 1% in *runt*, $P \leq 0.01$) (Fig. 2g). Embryos defective for *Toll-2* alone had intermediate defects, indicating that the other Toll receptors cannot fully substitute for Toll-2 in orienting edge formation. These results indicate that Toll receptors are required for rapid edge contraction and directional edge formation, suggesting that a common mechanism underlies both steps of cell rearrangement. Physical forces generated by the intercalation of subsets of cells can reinforce myosin polarity[10] and trigger passive cell stretching in neighbouring cells[20], perhaps allowing for substantial elongation in embryos that express a partial complement of Toll receptors.

Toll receptors and planar polarity

Cell intercalation in *Drosophila* is driven by the planar polarized activity of myosin II, which promotes the contraction of AP edges[6–10], and Par-3, which excludes myosin and stabilizes adhesion at DV edges[6,21]. To determine whether Toll receptors are required for myosin II and Par-3 localization, we used automated methods to analyse planar polarity at single-cell resolution[39]. In wild-type embryos, myosin II was enriched 1.30 ± 0.02-fold at AP edges and Par-3 was enriched 1.71 ± 0.03-fold at DV edges (Fig. 3a, e, f). By contrast, *Toll-2,6,8* mutants had a 47% reduction in myosin II planar polarity (1.16 ± 0.01, $P < 0.0001$) and a 48% reduction in Par-3 planar polarity (1.37 ± 0.02, $P < 0.0001$) (Fig. 3b, e, f). Similar defects were observed in *runt* mutants, although planar polarity was more severely affected in *eve* mutants (1.21 ± 0.01 for Par-3 and 1.09 ± 0.02 for myosin, $P < 0.0001$) (Fig. 3c–f and Extended Data Fig. 5). Toll receptor expression is reduced in *eve* mutants, whereas *runt* mutants have increased expression (Fig. 1g–l), suggesting that distinct mechanisms may underlie the defects in these two backgrounds. Apical–basal polarity was unaffected in Toll receptor mutants (Extended Data Fig. 3l), and planar polarity was not further reduced in *Toll-2,6,7,8* quadruple mutants (Extended Data Fig. 5g, h). These results demonstrate that Toll-2, Toll-6 and Toll-8 act together to regulate myosin II and Par-3 planar polarity.

Par-3 and myosin II planar polarity displayed regional differences in *Toll-2* mutants. Planar polarity occurred normally in Toll-8-expressing cells, most of which also express Toll-6, but was significantly reduced in Toll-8-negative cells, the majority of which do not express any Toll receptors (Fig. 3e, g, i and Extended Data Fig. 6a). Similarly, in *Toll-6,8* mutants, Toll-2-expressing cells had wild-type planar polarity, whereas cells that did not express any of these receptors had significant defects (Fig. 3h, j). Therefore, embryos expressing only one or two Toll receptors have localized planar polarity defects in the regions of missing receptor expression.

In *eve* mutants, which almost completely lack planar polarized myosin, residual myosin cables still form at the posterior boundaries of *Toll-2* stripes (Fig. 4d and Extended Data Fig. 6b, c), suggesting that differences in Toll receptor activity may induce planar polarity. To test this, we expressed Toll-2 and Toll-8 in stripes in the late embryo using the *engrailed–Gal4* driver. The anterior boundary of each *engrailed* stripe

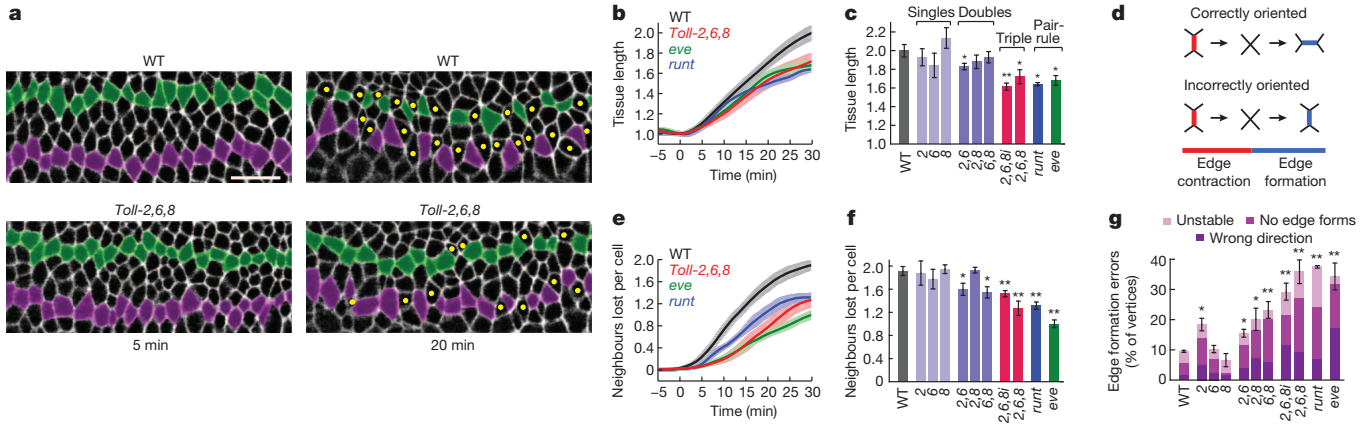


Figure 2 | Toll-2, Toll-6 and Toll-8 regulate cell intercalation and axis elongation. **a**, Stills from time-lapse movies of a wild-type (WT) embryo (top) and a *Toll-8* mutant injected with *Toll-2* and *Toll-6* dsRNAs (*Toll-2,6,8*) (bottom). Resille–GFP (white). $t = 0$, onset of elongation. In wild type, nearly all initially adjacent cells become separated by intercalated cells (yellow dots). In *Toll-2,6,8* embryos, many cells fail to separate. Anterior left, ventral down. Scale bar, 20 μm. **b**, **c**, Axis elongation (tissue AP length relative to $t = 0$) over time (**b**) and at 30 min (**c**). **d**, Edge contraction and formation. **e**, **f**, Cell rearrangements over time (**e**) and at 30 min (**f**). Single average values were obtained for each embryo; plots show the mean ± s.e.m. across embryos. **b**–**f**, $n = 3$–8 embryos per genotype, 164–365 cells per embryo (Supplementary Table 2). **g**, Edge formation errors. $n = 3$–9 embryos per genotype, 42–104 vertices per embryo. $^*P = 0.01$–0.03, $^{**}P < 0.005$ (unpaired *t*-test). WT (*Spider–GFP*); *2* (*Resille–GFP* + *Toll-2* dsRNA); *6* (*Resille–GFP* + *Toll-6* dsRNA); *8* (*Resille–GFP; Toll-8*$^{59/145}$); *2,6* (*Resille–GFP* + *Toll-2/Toll-6* dsRNAs); *2,8* (*Resille–GFP; Toll-8*$^{59/145}$ + *Toll-2* dsRNA); *6,8* (*Toll-2*$^{\Delta 76}$*/CyO; Toll-8*^{59}, *Toll-6*^{5A}, *Spider–GFP*); *2,6,8 i* (*Resille–GFP; Toll-8*$^{59/145}$+*Toll-2/Toll-6* dsRNAs); *2,6,8* (*Toll-2*$^{\Delta 76}$*; Toll-8*^{59}, *Toll-6*^{5A}, *Spider–GFP*); *runt* (*runt*^{LB5}; *Spider–GFP/+*); *eve* (*eve*^{R13}*; Spider–GFP/+*).

is situated anterior to the denticle-forming cells, in a region where myosin II is not strongly planar polarized (Fig. 4a). Ectopic Toll-2 or Toll-8 led to a strong recruitment of myosin II to the anterior boundary of the *engrailed* domain (Fig. 4b, c, e and Supplementary Videos 3–5) and increased contractile activity at this boundary, as measured by laser ablation (Fig. 4f). These results demonstrate that local differences in Toll-2 or Toll-8 expression are sufficient to induce myosin planar polarity *in vivo*.

Heterophilic Toll receptor interactions

Drosophila Toll receptors are known to bind to Spätzle/DNT neurotrophin-related growth factors[23,28,40,41], but the ligands detected by Toll receptors during convergent extension are not known. In one model, heterophilic interactions between receptors expressed on adjacent stripes of cells could induce actomyosin contractility at AP cell edges. Alternatively, homophilic interactions between receptors expressed in the same stripe could suppress actomyosin contractility and stabilize adhesion at DV edges. To investigate these possibilities, we tested for interactions between Toll receptors in *Drosophila* S2R+ cells[42]. Cells expressing Toll-2 displayed increased affinity for a soluble, pentamerized form of the Toll-8 extracellular domain (Fig. 5a, b). By contrast, cells expressing Toll-2 displayed decreased affinity for the Toll-2 extracellular domain (Fig. 5c, d). These results indicate that Toll-2 and Toll-8 can interact in a heterophilic manner in cultured cells.

To test whether Toll receptors can promote interactions between cells, we performed cell-mixing experiments. *Drosophila* S2R+ cells

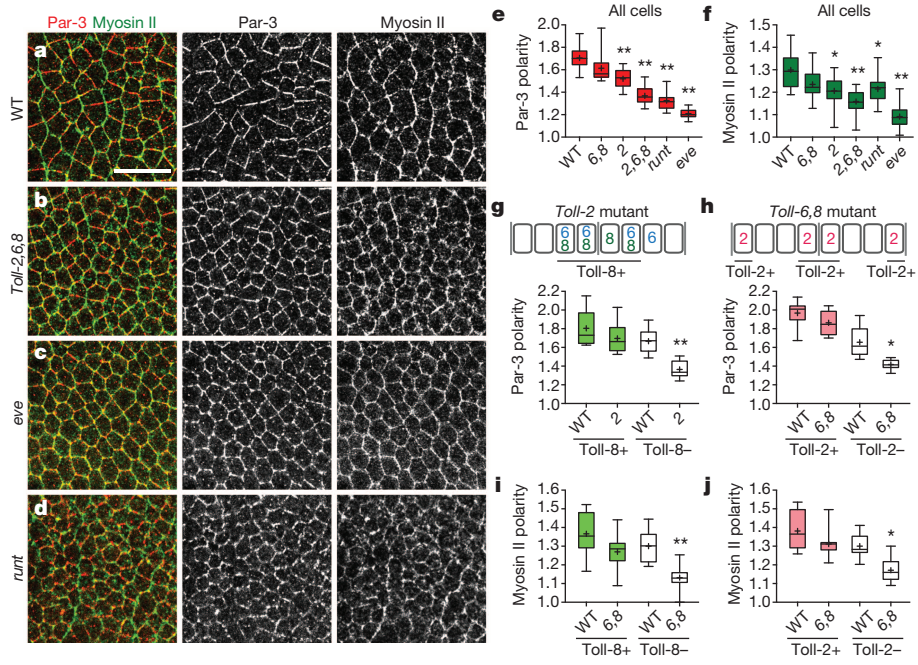


Figure 3 | Toll receptors are required for myosin II and Par-3 planar polarity. **a**–**d**, Stage 7 wild-type (**a**), *Toll-2,6,8* (**b**), *eve* (**c**) and *runt* (**d**) embryos. Par-3 (red, middle), myosin II (green, right). **e**–**j**, Par-3 and myosin II planar polarity in all cells (**e**, **f**) and subsets of cells (**g**–**j**). Horizontal line, median; +, mean; boxes, second and third quartiles; whiskers, 95% confidence interval. Single average values were obtained for each embryo; plots show the distribution of values across embryos. $^*P \leq 0.005$, $^{**}P < 0.0001$ (unpaired *t*-test). $n = 11$–19 embryos per genotype; 2,445–4,698 cells per embryo (Supplementary Table 2). Anterior left, ventral down. Scale bar, 20 μm.

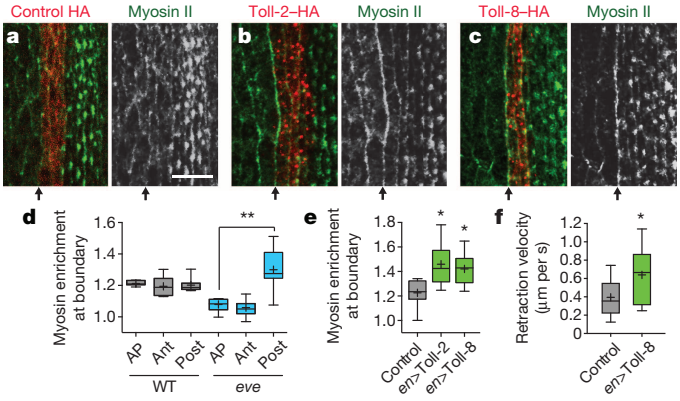


Figure 4 | Myosin II localization and activity are enhanced at boundaries of Toll-2 and Toll-8 expression. **a–c,** Stage 15 embryos expressing control β-catenin–HA (**a**), Toll-2–HA (**b**), or Toll-8–HA (**c**) expressed with *engrailed–Gal4*. Myosin II (green, white), HA (red). Arrows, anterior boundary of the *engrailed* domain. Ventral views. Scale bar, 10 μm. **d**, Myosin levels are increased at the posterior boundary of *Toll-2* stripes in *eve* mutants ($P = 0.00001$). All edges oriented 75–90° relative to the AP axis (AP) or edges only at anterior (Ant) or posterior (Post) boundaries of *Toll-2* stripes; edge values were normalized to average edge intensity. **e**, Myosin levels are increased at the anterior boundary of ectopic Toll-2 and Toll-8 expression. **f**, Peak retraction velocities following laser ablation are increased at the anterior boundary of ectopic Toll-8 expression. Horizontal line, median; boxes, second and third quartiles; whiskers, 95% confidence interval. Single average values were obtained for each embryo; plots show the distribution of values across embryos. $*P \leq 0.008$, $**P \leq 0.0001$ (unpaired *t*-test). **d**, **e**, $n = 6–15$ embryos per genotype, **f**, $n = 16–17$ ablations per genotype (Supplementary Table 2).

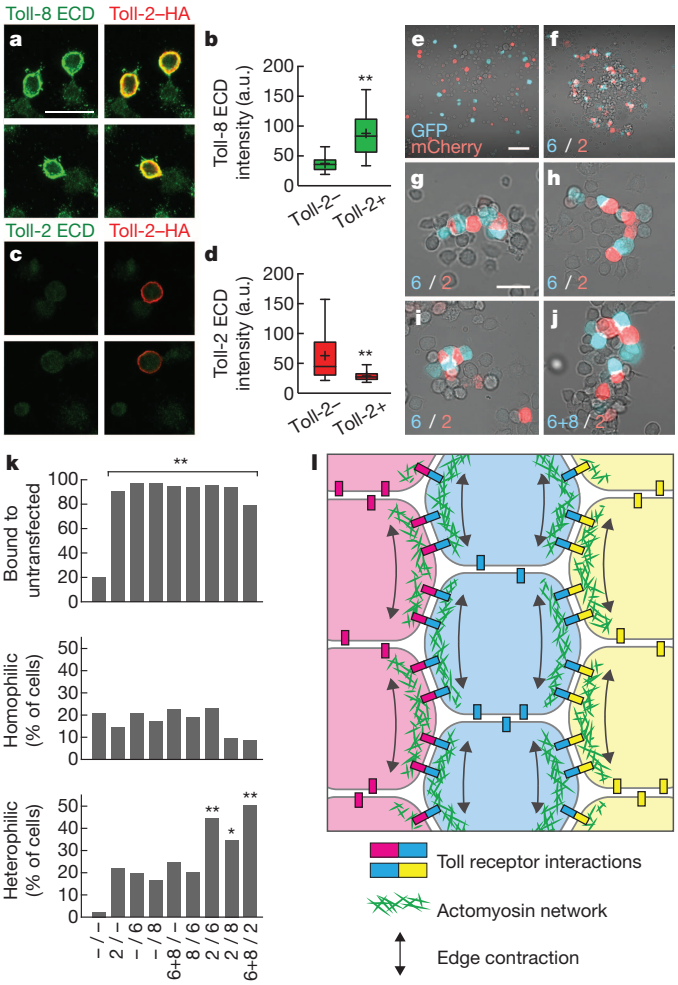


Figure 5 | Toll receptors mediate heterophilic interactions between cells. **a–d**, *Drosophila* S2R+ cells expressing Toll-2–HA (red) incubated with pentamerized Toll-8 (**a**, **b**) or Toll-2 (**c**, **d**) extracellular domains (ECD) (green). Toll-8 ECD bound more strongly (**b**) and Toll-2 ECD bound less strongly (**d**) to Toll-2-positive (Toll-2+) cells compared with Toll-2-negative (Toll-2−) cells ($P < 0.00001$, unpaired *t*-test). Horizontal line, median; boxes, second and third quartiles; whiskers, 95% confidence interval. **e–k**, Interactions between cells expressing myosin–GFP (cyan, sample listed before the / symbol) or myosin–mCherry (red, sample listed after the / symbol) with the indicated Toll receptors (−, myosin marker alone). Receptor-expressing cells displayed increased binding to untransfected cells ($P \leq 0.0001$, Chi-square test). Heterophilic binding was increased between cells expressing Toll-2 and Toll-6 ($P \leq 0.0003$), Toll-2 and Toll-8 ($P < 0.05$), and Toll-2 and Toll-6 + Toll-8 ($P \leq 0.0001$) (Chi-square test). $*P = 0.01–0.05$, $**P \leq 0.0003$. **l**, Model showing heterophilic interactions between Toll receptors recruit myosin II, promoting oriented cell rearrangements and convergent extension. **b**, **d**, $n = 170–176$ cells per condition, **k**, $n = 85–123$ transfected cells per condition (Supplementary Table 2). Scale bars, 20 μm (**a**, **c**, **g–j**), 100 μm (**e**, **f**).

normally do not aggregate, but cells expressing Toll-2, Toll-6, or Toll-8 aggregated with untransfected cells at high frequency, indicating that these receptors can bind to proteins present on S2R+ cells (Fig. 5e, f, k)[31,32]. Homophilic interactions between cells were not enhanced by Toll receptor expression (Fig. 5k). By contrast, Toll-2-positive cells formed extensive heterophilic contacts with cells expressing Toll-6 and/or Toll-8, creating chains of cells expressing alternating Toll receptors (Fig. 5g–k). Heterophilic interactions were not observed between cells expressing Toll-6 and Toll-8, which are often coexpressed within the same stripes (Fig. 1p). These results indicate that Toll-2 can promote heterophilic interactions with cells expressing Toll-6 or Toll-8. Embryos expressing any one receptor still display significant planar polarity and intercalary behaviour, suggesting that these proteins also interact with additional binding partners to generate planar polarity.

Discussion

Together, these results demonstrate that the spatial signals that establish planar polarity and direct polarized cell behaviour during convergent extension in *Drosophila* are encoded at the cell surface by three Toll family receptors expressed in overlapping stripes along the AP axis of the embryo. Simultaneous disruption of Toll-2, Toll-6 and Toll-8 significantly impairs planar polarity, cell intercalation, and convergent extension, and removing one or two receptors disrupts planar polarity in distinct subsets of cells, indicating that these proteins serve non-redundant and highly localized functions. These findings support a model in which planar polarity is induced by interactions between neighbouring cells with different levels of Toll receptor activity (Fig. 5l). Therefore, *Drosophila* Toll receptors provide the basis of a spatial code that translates patterned Eve and Runt transcriptional activity into planar polarized actomyosin contractility, linking positional information provided by the embryonic AP patterning system to the essential cell behaviours that drive convergent extension. The Toll receptor code is incomplete in certain regions, such as the parasegmental boundaries, suggesting the existence of additional polarity cues at these interfaces. *Toll-2,6,8* mutants are similar to *runt* mutants with respect to all measures of cell rearrangement and planar polarity, but are not as severe as *eve* mutants. Thus, although *Toll-2,6,8* mutants recapitulate much of the *eve* mutant phenotype, Eve likely has additional targets important for planar polarity.

Toll family receptors have a highly conserved structure in vertebrates and invertebrates, including extracellular LRR motifs that are often present in proteins involved in cell adhesion and cell–cell recognition[43]. Although individual receptors are not orthologous between flies and humans[25], mammalian Toll-like receptors are required for epithelial regeneration and wound healing, processes that involve dynamic and spatially regulated changes in cell adhesion[44–46]. In the innate immune system, pathogen detection by Toll family receptors activates transcriptional pathways mediated by NF-κB and MAP kinase signalling[23–26]. However, the spatial information provided by patterned Toll receptor

expression in *Drosophila,* as well as the rapid timescale of cell rearrangements during convergent extension, suggest a more direct connection between Toll receptor signalling and the cellular contractile machinery. Consistent with this possibility, activation of mammalian Toll-like receptors in dendritic cells induces a rapid remodelling of the actin cytoskeleton[47] and mammalian Toll-like receptors can inhibit neurite outgrowth and trigger rapid growth cone collapse in neurons[48,49], reminiscent of Toll receptor functions in the *Drosophila* nervous system[40,41,50]. Elucidating the mechanisms that link Toll family receptors to dynamic changes in cell polarity and behaviour may provide insight into conserved and relatively unexplored aspects of Toll receptor signalling.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 May; accepted 9 October 2014.

Published online 2 November 2014.

1. Keller, R. *et al.* Mechanisms of convergence and extension by cell intercalation. *Phil. Trans. R. Soc. Lond. B* **355,** 897–922 (2000).
2. Zallen, J. A. Planar polarity and tissue morphogenesis. *Cell* **129,** 1051–1063 (2007).
3. Wallingford, J. B. Planar cell polarity and the developmental control of cell behavior in vertebrate embryos. *Annu. Rev. Cell Dev. Biol.* **28,** 627–653 (2012).
4. Solnica-Krezel, L. & Sepich, D. S. Gastrulation: making and shaping germ layers. *Annu. Rev. Cell Dev. Biol.* **28,** 687–717 (2012).
5. Walck-Shannon, E. & Hardin, J. Cell intercalation from top to bottom. *Nature Rev. Mol. Cell Biol.* **15,** 34–48 (2014).
6. Zallen, J. A. & Wieschaus, E. Patterned gene expression directs bipolar planar polarity in *Drosophila. Dev. Cell* **6,** 343–355 (2004).
7. Bertet, C., Sulak, L. & Lecuit, T. Myosin-dependent junction remodelling controls planar cell intercalation and axis elongation. *Nature* **429,** 667–671 (2004).
8. Blankenship, J. T., Backovic, S. T., Sanny, J. S. P., Weitz, O. & Zallen, J. A. Multicellular rosette formation links planar cell polarity to tissue morphogenesis. *Dev. Cell* **11,** 459–470 (2006).
9. Rauzi, M., Verant, P., Lecuit, T. & Lenne, P.-F. Nature and anisotropy of cortical forces orienting *Drosophila* tissue morphogenesis. *Nature Cell Biol.* **10,** 1401–1410 (2008).
10. Fernández-González, R., Simões, S. de M., Röper, J.-C., Eaton, S. & Zallen, J. A. Myosin II dynamics are regulated by tension in intercalating cells. *Dev. Cell* **17,** 736–743 (2009).
11. Nishimura, T. & Takeichi, M. Shroom3-mediated recruitment of Rho kinases to the apical cell junctions regulates epithelial and neuroepithelial planar remodeling. *Development* **135,** 1493–1502 (2008).
12. Nishimura, T., Honda, H. & Takeichi, M. Planar cell polarity links axes of spatial dynamics in neural-tube closure. *Cell* **149,** 1084–1097 (2012).
13. Lienkamp, S. S. *et al.* Vertebrate kidney tubules elongate using a planar cell polarity-dependent, rosette-based mechanism of convergent extension. *Nature Genet.* **44,** 1382–1387 (2012).
14. Mahaffey, J. P., Grego-Bessa, J., Liem, K. F. & Anderson, K. V. Cofilin and Vangl2 cooperate in the initiation of planar cell polarity in the mouse embryo. *Development* **140,** 1262–1271 (2013).
15. Shindo, A. & Wallingford, J. B. PCP and septins compartmentalize cortical actomyosin to direct collective cell movement. *Science* **343,** 649–652 (2014).
16. Williams, M., Yen, W., Lu, X. & Sutherland, A. Distinct apical and basolateral mechanisms drive planar cell polarity-dependent convergent extension of the mouse neural plate. *Dev. Cell* **29,** 34–46 (2014).
17. Irvine, K. D. & Wieschaus, E. Cell intercalation during *Drosophila* germband extension and its regulation by pair-rule segmentation genes. *Development* **120,** 827–841 (1994).
18. Ninomiya, H., Elinson, R. P. & Winklbauer, R. Antero-posterior tissue polarity links mesoderm convergent extension to axial patterning. *Nature* **430,** 364–367 (2004).
19. St Johnston, D. & Nüsslein-Volhard, C. The origin of pattern and polarity in the *Drosophila* embryo. *Cell* **68,** 201–219 (1992).
20. Butler, L. C. *et al.* Cell shape changes indicate a role for extrinsic tensile forces in *Drosophila* germ-band extension. *Nature Cell Biol.* **11,** 859–864 (2009).
21. Simões, S. de M. *et al.* Rho-kinase directs Bazooka/Par-3 planar polarity during *Drosophila* axis elongation. *Dev. Cell* **19,** 377–388 (2010).
22. Wieschaus, E., Sweeton, D. & Costa, M. in *Gastrulation* 213–223 (Springer, 1992).
23. Brennan, C. A. & Anderson, K. V. *Drosophila*: the genetics of innate immune recognition and response. *Annu. Rev. Immunol.* **22,** 457–483 (2004).
24. Janeway, C. A. & Medzhitov, R. Innate immune recognition. *Annu. Rev. Immunol.* **20,** 197–216 (2002).
25. Leulier, F. & Lemaitre, B. Toll-like receptors—taking an evolutionary approach. *Nature Rev. Genet.* **9,** 165–178 (2008).
26. Kawai, T. & Akira, S. The role of pattern-recognition receptors in innate immunity: update on Toll-like receptors. *Nature Immunol.* **11,** 373–384 (2010).
27. Tauszig, S., Jouanguy, E., Hoffmann, J. A. & Imler, J. L. Toll-related receptors and the control of antimicrobial peptide expression in *Drosophila. Proc. Natl Acad. Sci. USA* **97,** 10520–10525 (2000).
28. Morisato, D. & Anderson, K. V. Signaling pathways that establish the dorsal-ventral pattern of the *Drosophila* embryo. *Annu. Rev. Genet.* **29,** 371–399 (1995).
29. Chiang, C. & Beachy, P. A. Expression of a novel *Toll*-like gene spans the parasegment boundary and contributes to *hedgehog* function in the adult eye of *Drosophila. Mech. Dev.* **47,** 225–239 (1994).
30. Kambris, Z., Hoffmann, J. A., Imler, J.-L. & Capovilla, M. Tissue and stage-specific expression of the *Tolls* in *Drosophila* embryos. *Gene Expr. Patterns* **2,** 311–317 (2002).
31. Eldon, E. *et al.* The *Drosophila 18 wheeler* is required for morphogenesis and has striking similarities to *Toll. Development* **120,** 885–899 (1994).
32. Keith, F. J. & Gay, N. J. The *Drosophila* membrane receptor Toll can function to promote cellular adhesion. *EMBO J.* **9,** 4299–4306 (1990).
33. Kim, S., Chung, S., Yoon, J., Choi, K.-W. & Yim, J. Ectopic expression of Tollo/Toll-8 antagonizes Dpp signaling and induces cell sorting in the *Drosophila* wing. *Genesis* **44,** 541–549 (2006).
34. Kleve, C. D., Siler, D. A., Syed, S. K. & Eldon, E. D. Expression of *18-wheeler* in the follicle cell epithelium affects cell migration and egg morphology in *Drosophila. Dev. Dyn.* **235,** 1953–1961 (2006).
35. Kolesnikov, T. & Beckendorf, S. K. 18 wheeler regulates apical constriction of salivary gland cells via the Rho-GTPase-signaling pathway. *Dev. Biol.* **307,** 53–61 (2007).
36. Paré, A. *et al.* Visualization of individual *Scr* mRNAs during *Drosophila* embryogenesis yields evidence for transcriptional bursting. *Curr. Biol.* **19,** 2037–2042 (2009).
37. Cermak, T. *et al.* Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res.* **39,** e82 (2011).
38. Tamada, M., Farrell, D. L. & Zallen, J. A. Abl regulates planar polarized junctional dynamics through β-catenin tyrosine phosphorylation. *Dev. Cell* **22,** 309–319 (2012).
39. Kasza, K. E., Farrell, D. L. & Zallen, J. A. Spatiotemporal control of epithelial remodeling by regulated myosin phosphorylation. *Proc. Natl Acad. Sci. USA* **111,** 11732–11737 (2014).
40. McIlroy, G. *et al.* Toll-6 and Toll-7 function as neurotrophin receptors in the *Drosophila melanogaster* CNS. *Nature Neurosci.* **16,** 1248–1256 (2013).
41. Ballard, S. L., Miller, D. L. & Ganetzky, B. Retrograde neurotrophin signaling through Tollo regulates synaptic growth in *Drosophila. J. Cell Biol.* **204,** 1157–1172 (2014).
42. Özkan, E. *et al.* An extracellular interactome of immunoglobulin and LRR proteins reveals receptor–ligand networks. *Cell* **154,** 228–239 (2013).
43. de Wit, J., Hong, W., Luo, L. & Ghosh, A. Role of leucine-rich repeat proteins in the development and function of neural circuits. *Annu. Rev. Cell Dev. Biol.* **27,** 697–729 (2011).
44. Rakoff-Nahoum, S. & Medzhitov, R. Toll-like receptors and cancer. *Nature Rev. Cancer* **9,** 57–63 (2009).
45. Grote, K., Schütt, H. & Schieffer, B. Toll-like receptors in angiogenesis. *Scientific World J.* **11,** 981–991 (2011).
46. Huebener, P. & Schwabe, R. F. Regulation of wound healing and organ fibrosis by toll-like receptors. *Biochim. Biophy. Acta* **1832,** 1005–1017 (2013).
47. West, M. A. *et al.* Enhanced dendritic cell antigen capture via toll-like receptor-induced actin remodeling. *Science* **305,** 1153–1157 (2004).
48. Ma, Y. *et al.* Toll-like receptor 8 functions as a negative regulator of neurite outgrowth and inducer of neuronal apoptosis. *J. Cell Biol.* **175,** 209–215 (2006).
49. Cameron, J. S. *et al.* Toll-like receptor 3 is a potent negative regulator of axonal growth in mammals. *J. Neurosci.* **27,** 13033–13041 (2007).
50. Rose, D. *et al.* Toll, a muscle cell surface molecule, locally inhibits synaptic initiation of the RP3 motoneuron growth cone in *Drosophila. Development* **124,** 1561–1571 (1997).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank K. Anderson, K. Kasza, W. Razzell, G. Sabio, M. Shirasu-Hiza, A. Spencer, M. Tamada and R. Zallen for comments on the manuscript, B. Glick for the fast-folding YFP, M. Buszczak for pUASp-w-attB, and the BAC-Recombineering Core Facility at the University of Chicago for Toll-8–YFP. This work was funded by NIH/NIGMS grants GM079340 and GM102803 to J.A.Z. J.A.Z. is an Early Career Scientist of the Howard Hughes Medical Institute.

Author Contributions A.C.P., A.V. and J.A.Z. designed the study. A.C.P., A.V., C.T.F. and Z.M. performed the experiments, D.L.F. and A.M. performed the computational analysis, and A.C.P. and J.A.Z. wrote the manuscript. All authors participated in analysis of the data and in producing the final version of the manuscript.

Author Information The complete RNA sequencing data set is available on the Gene Expression Omnibus, accession code GSE61689. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.A.Z. (zallenj@mskcc.org).

LETTER

doi:10.1038/nature13914

A dust-parallax distance of 19 megaparsecs to the supermassive black hole in NGC 4151

Sebastian F. Hönig[1,2], Darach Watson[1], Makoto Kishimoto[3] & Jens Hjorth[1]

The active galaxy NGC 4151 has a crucial role as one of only two active galactic nuclei for which black hole mass measurements based on emission line reverberation mapping can be calibrated against other dynamical techniques[1–3]. Unfortunately, effective calibration requires accurate knowledge of the distance to NGC 4151, which is not at present available[4]. Recently reported distances range from 4 to 29 megaparsecs[5–7]. Strong peculiar motions make a redshift-based distance very uncertain, and the geometry of the galaxy and its nucleus prohibit accurate measurements using other techniques. Here we report a dust-parallax distance to NGC 4151 of $19.0^{+2.4}_{-2.6}$ megaparsecs. The measurement is based on an adaptation of a geometric method that uses the emission line regions of active galaxies[8]. Because these regions are too small to be imaged with present technology, we use instead the ratio of the physical and angular sizes of the more extended hot-dust emission[9] as determined from time delays[10] and infrared interferometry[11–14]. This distance leads to an approximately 1.4-fold increase in the dynamical black hole mass, implying a corresponding correction to emission line reverberation masses of black holes if they are calibrated against the two objects with additional dynamical masses.

The central black hole in an active galactic nucleus (AGN) is surrounded by a putative accretion disk that emits predominantly in ultraviolet and optical wavelengths. At large distances from this central emission source, the gas is cool enough for dust to survive (temperature $\lesssim$ 1,500 K). This 'dusty torus' absorbs the ultraviolet–optical radiation and thermally reemits the energy in the infrared. Thus, any variability in the ultraviolet–optical emission will be detected in the dust emission with some time delay. Near-infrared reverberation mapping measures the time lag, τ, between the ultraviolet–optical variability and the corresponding changes in emission of the hot dust. The hottest dust is located at about the sublimation radius, $R_{\rm sub} = R(T \approx 1{,}500\,{\rm K})$. The time lag can be converted into a physical size using $R_\tau = \tau c$, where c is the speed of light. Typically, time lags are in the range of several tens to hundreds of days, which corresponds to physical sizes of the order of 0.1 pc, with a square-root dependence on luminosity[12,15].

In parallel, infrared interferometry at the same wavelength measures the angular size, ρ, of the same emission region. The angular and physical sizes are trigonometrically related by $\sin(\rho) = R_\tau/D_{\rm A}$, where $D_{\rm A}$ is the angular-diameter distance to the object. For small angles, $\sin(\rho) \approx \rho$ and, accounting for cosmological time dilation, we obtain $D_{\rm A}({\rm Mpc}) = 0.173\tau({\rm days})/(\rho({\rm mas})(1+z))$, which forms the basis of the distance measurement presented here (Fig. 1). A geometric technique was first proposed for broad emission lines[8]. Unfortunately, the typical angular size of the broad-emission regions of bright AGNs is of the order of 0.001–0.01 mas, which is too small to be spatially resolved with today's optical long-baseline interferometers. The dust continuum emission, however, is larger by a factor of ~4, and infrared interferometers have now managed to resolve about a dozen AGNs[11,12,14,16]. Moreover, using dust emission requires only photometric reverberation mapping instead of the spectral resolution of emission lines. Finally, dust physics is arguably easier to model than gas line emission.

To determine the distance to the supermassive black hole in NGC 4151, we make use of interferometry obtained with the two Keck telescopes and monitoring data from the literature. V-(wavelength 0.55 μm) and K-band (2.2 μm) photometric monitoring from 2001 to 2006[10] traces the ultraviolet–optical and hot-dust emission, respectively. Because long-term brightness changes can cause τ and ρ to increase or decrease, monitoring and interferometry data should be recorded more or less contemporaneously. We use six Keck interferometry measurements made between 2003 and 2010[11–14]. These overlap with the monitoring data, and inspection of long-term brightness trends showed that fluctuations were moderate between 2000 and 2010[16]. Indeed, no significant change in size has been detected for this set of interferometry and variability data[16,17].

When comparing angular and physical sizes, it is important to make sure that they refer to the same physical region. First, observations of the dust emission are preferably made at the same waveband (here the K band). Second, the spatial distribution of the dust around the AGN affects the observed sizes (Fig. 1): dust that is homogeneously distributed will result in a larger apparent size than will a compact dusty region,

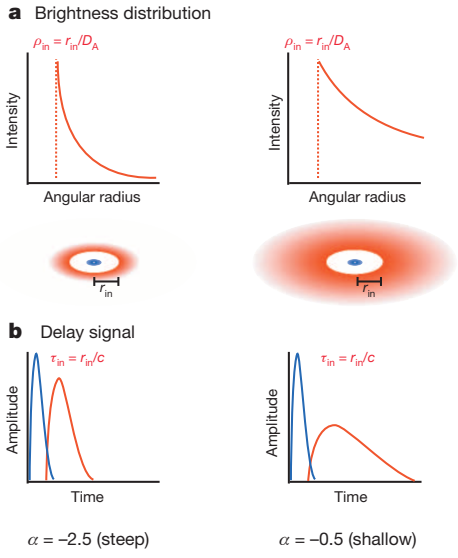


Figure 1 | Effect of the brightness distribution on the observed sizes and time lags. The bottom row (**b**) illustrates effects on the time lag signal of varying the brightness distribution (blue, optical; red, near-infrared), and the upper row (**a**) outlines the corresponding (interferometric) brightness distribution in the near-infrared. Compact distributions lead to shorter time lags and smaller interferometric radii than do shallow profiles. This information is encoded in the shape (width, amplitude) of the light curve. With a simple power-law parameterization, this smearing effect can be accounted for to determine the time lag and angular size of the innermost radius of the brightness distribution. The simultaneous modelling of light curves and interferometry results in a very precise angular distance measurement.

[1]Dark Cosmology Centre, Niels Bohr Institute, University of Copenhagen, Juliane Maries Vej 30, 2100 Copenhagen Ø, Denmark. [2]School of Physics & Astronomy, University of Southampton, Southampton SO17 1BJ, UK. [3]Department of Physics, Faculty of Science, Kyoto Sangyo University, Kamigamo-motoyama, Kita-ku, Kyoto 603-8555, Japan.

because for the former a larger region contributes to the emission at a given wavelength. As a consequence, the variability signal in the *K* band will show some degree of smoothing with respect to the *V* band signal, depending on the distribution. This also involves a shift of the peak time lag. At the same time, the size measured by interferometry will appear correspondingly smaller or larger. This distribution effect in both types of data can be effectively modelled by means of a disk model[17,18], assuming that the dust is heated by AGN radiation and the projected brightness distribution is represented by the power law $S(r) \propto r^{\alpha}$ (Methods). This geometry is in line with theoretical expectations and observational evidence of the hot-dust region[19,20]. Indeed, the model has been successfully applied to reproduce multiwavelength, multi-baseline interferometry of several AGNs[14,21] (including NGC 4151), as well as the light curve of NGC 4151[17,22].

The common reference size in such a model is the inner radius, $r_{\rm in}$, of the brightness distribution. For reverberation mapping and interferometry, this corresponds to a reference time lag $\tau_{\rm in} = r_{\rm in}/c$ and angular size $\rho_{\rm in} \approx r_{\rm in}/D_{\rm A}$ of the inner boundary of the brightness distribution. Other parameters that may influence the observationally inferred physical and angular size of $r_{\rm in}$ are the disk geometry of the emission region and the dust properties. For the inclination and disk orientation, we use observational constraints based on the dynamics of the emission line region in NGC 4151 and polarimetry[23–25]. We do not consider the radio jet in this study[26], because the available data do not allow both position angle and inclination to be reliably established at the same time. The absorption efficiency of the dust is implicitly included in our parameterization of the brightness distribution. Moreover, the sublimation temperature does not affect the distance determination because it scales in the same way for the reference angular sizes and the reference time lags (Methods).

Because the light curves are sampled with finite and varying gaps between the observations, we simulated 1,250 random, continuous representations of the data using the AGN variability pattern derived from the structure function. We calculated $\tau_{\rm in}$, α and $\rho_{\rm in}$ simultaneously, given the observationally constrained inclination and disk orientation as priors. This resulted in 1,250 estimates of $D_{\rm A}$, which are shown in Fig. 2. An important feature of this process is that, although determining the reference time lag or the reference angular size individually is quite uncertain, both parameters are strongly correlated with the dust brightness distribution. Thus, the ratio, that is, $D_{\rm A}$, can be constrained with much higher precision than can the reference time lag or reference angular size individually, if $\tau_{\rm in}$ and $\rho_{\rm in}$ are calculated simultaneously given the inferred α.

We obtain an angular-diameter distance to NGC 4151 of $D_{\rm A} = 19.0^{+2.4}_{-2.6}$ Mpc (Fig. 2, inset probability distribution). The error bars include statistical uncertainties from the reverberation and interferometric observations, as well as the systematic uncertainties introduced by the geometry, the brightness distribution and the uncertainty in the contributions of the host and the putative accretion disk to the K-band interferometry. These uncertainties have been accounted for in Monte Carlo simulations when sampling the data (Methods). The new distance clarifies the situation for NGC 4151. The galaxy is in the vicinity of the Virgo cluster ($<30^{\circ}$ angular separation from the Virgo cluster centre), resulting in strong peculiar motion with respect to the Hubble flow[6,27]. Therefore, any recession-velocity-dependent distance has to be considered uncertain. Geometric megamaser distances require that the nuclear region is seen very close to edge-on, which is generally not the case for unobscured AGN. Moreover, a direct distance estimate based on the Tully–Fisher relation is difficult because of the face-on view of the galactic disk, which makes it difficult to determine the required rotational velocities. Attempts to estimate the distance this way resulted in a wide range of values between ~4 and 20 Mpc (refs 5, 6). More recently, a luminosity distance of 29.2 ± 0.5 Mpc was suggested on the basis of near-infrared reverberation mapping only and a model for the absorption and re-emission of dust, but cannot be reconciled with our new result or the other estimates[7].

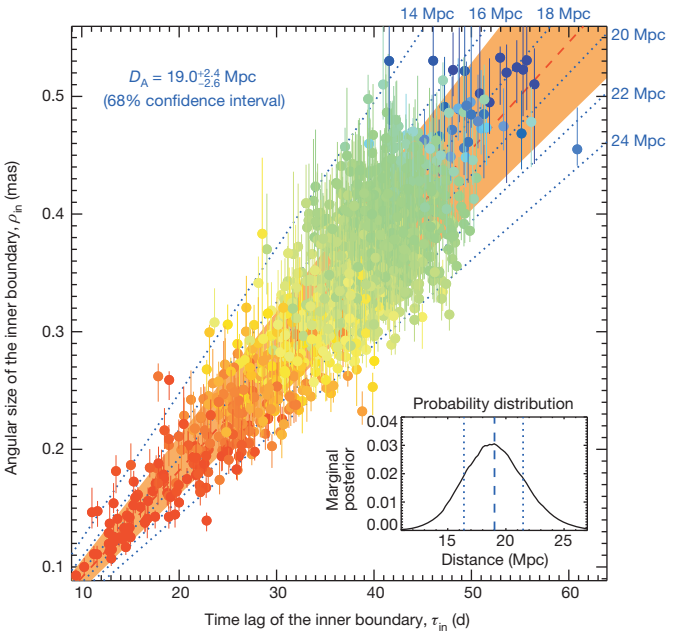


Figure 2 | Relating time lags and angular sizes to measure the absolute distance to NGC 4151. The coloured circles show the modelled reference time lags, $\tau_{\rm in}$, and associated angular sizes, $\rho_{\rm in}$ (1,250 random realizations of the *V*-band light curve; 68% confidence interval shown as error bars). The dotted blue lines mark $\tau_{\rm in}/\rho_{\rm in}$ ratios corresponding to distances in the range $D_{\rm A}$ = 14–24 Mpc. The distribution median and 68% confidence interval are marked by the dashed red line and the orange shaded area. Points are colour-coded to indicate the brightness distribution power-law index α, in the range $\alpha = 2$ (red; shallow) to $\alpha = -15$ (blue; steep). Inset, probability distribution function of the inferred distance, with mean and 68% confidence interval indicated by the dashed and dotted lines, respectively.

The new precise distance to NGC 4151 is relevant because this galaxy is a cornerstone in calibrating black hole masses inferred by different methods[2]; apart from NGC 3227, it is the only galaxy with a suitable mass estimated from reverberation mapping, stellar and gas dynamics. Of both galaxies, NGC 4151 has much better mass constraints[2], such that any systematic offset in distance will almost equally affect the calibration of black hole masses. Dynamical mass estimates relate the rotational velocity field of stars or gas surrounding the black hole to their distances from the AGN. In the process, observed angular distances have to be converted into physical distances, which requires knowledge of the absolute distance to the galaxy. The most recent mass estimates assume that $D_{\rm A}$ = 13.2 Mpc (refs 4, 28). A stellar-velocity-based mass was reported as $M_{\rm BH}^{\rm SD} = (3.76 \pm 1.15) \times 10^7 M_{\odot}$ (ref. 4), where $M_{\odot}$ is the solar mass. Our new measurement implies that this mass is underestimated by a factor of ~1.4, leading to a revised mass of $M_{\rm BH}^{\rm SD} = (5.4 \pm 1.8) \times 10^7 M_{\odot}$. Similarly, the correction to the distance increases the gas dynamical mass[28] from $M_{\rm BH}^{\rm GD} = 3.0^{+0.75}_{-2.2} \times 10^7 M_{\odot}$ to $M_{\rm BH}^{\rm GD} = 4.3^{+1.2}_{-3.2} \times 10^7 M_{\odot}$.

The new distance and the corrected values of $M_{\rm BH}^{\rm SD}$ and $M_{\rm BH}^{\rm GD}$ also affect the correction factor *f* that has to be invoked when converting reverberation time lags and velocities into black hole masses. The most recent reference value is $f = 4.31 \pm 1.05$, inferred from comparing reverberation mapping masses with black hole masses determined from the established relation between black hole mass $M_{\rm BH}$ and bulge stellar velocity dispersion σ_* (ref. 29). By using our corrected values for the dynamical black hole masses to calibrate the reverberation data[1], we find a range of $f = 5.2$–6.5 (reflecting the difference between gas and stellar dynamical masses), implying a systematic shift to larger masses. Such larger *f* values may be generally applicable, as also suggested by complex modelling of velocity-resolved reverberation mapping data[30].

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 26 June; accepted 3 October 2014.

1. Bentz, M. C. *et al.* A reverberation-based mass for the central black hole in NGC 4151. *Astrophys. J.* **651,** 775–781 (2006).
2. Peterson, B. M. Measuring the mass of supermassive black holes. *Space Sci. Rev.* **183,** 253–275 (2014).
3. Woo, J.-H. *et al.* Do quiescent and active galaxies have different $M_{BH}-\sigma_*$ relations? *Astrophys. J.* **772,** 49 (2013).
4. Onken, C. A. *et al.* The black hole mass of NGC 4151. II. Stellar dynamical measurement from near-infrared integral field spectroscopy. *Astrophys. J.* **791,** 37 (2014).
5. Theureau, G. *et al.* Kinematics of the local universe. XIII. 21-cm line measurements of 452 galaxies with the Nancay radiotelescope, JHK Tully-Fisher relation, and preliminary maps of the peculiar velocity field. *Astron. Astrophys.* **465,** 71–85 (2007).
6. NASA/IPAC Extragalactic Database, http://ned.ipac.caltech.edu (2014).
7. Yoshii, Y. *et al.* A new method for measuring extragalactic distances. *Astrophys. J.* **784,** L11 (2014).
8. Elvis, M. & Karovska, M. Quasar parallax: a method for determining direct geometrical distances to quasars. *Astrophys. J.* **581,** L67 (2002).
9. Hönig, S. F. Dust reverberation mapping in the era of big optical surveys and its cosmological application. *Astrophys. J.* **784,** L4 (2014).
10. Koshida, S. *et al.* Variation of inner radius of dust torus in NGC4151. *Astrophys. J.* **700,** L109 (2009).
11. Swain, M. *et al.* Interferometer observations of subparsec-scale infrared emission in the nucleus of NGC 4151. *Astrophys. J.* **596,** L163 (2003).
12. Kishimoto, M. *et al.* Exploring the inner region of type 1 AGNs with the Keck interferometer. *Astron. Astrophys.* **507,** L57–L60 (2009).
13. Pott, J.-U. *et al.* Luminosity-variation independent location of the circumnuclear hot dust in NGC 4151. *Astrophys. J.* **715,** 736–742 (2010).
14. Kishimoto, M. *et al.* The innermost dusty structure in active galactic nuclei as probed by the Keck interferometer. *Astron. Astrophys.* **527,** A121 (2011).
15. Suganuma, M. *et al.* Reverberation measurements of the inner radius of the dust torus in nearby Seyfert 1 galaxies. *Astrophys. J.* **639,** 46–63 (2006).
16. Kishimoto, M. *et al.* Evidence for a receding dust sublimation region around a supermassive black hole. *Astrophys. J.* **775,** L36 (2013).
17. Hönig, S. F. & Kishimoto, M. Constraining properties of dusty environments by infrared variability. *Astron. Astrophys.* **534,** A121 (2011).
18. Kishimoto, M. *et al.* Mapping the radial structure of AGN tori. *Astron. Astrophys.* **536,** A78 (2011).
19. Kawaguchi, T. & Mori, M. Near-infrared reverberation by dusty clumpy tori in active galactic nuclei. *Astrophys. J.* **737,** 105 (2011).
20. Landt, H. *et al.* The near-infrared broad emission line region of active galactic nuclei - II. The 1-μm continuum. *Mon. Not. R. Astron. Soc.* **414,** 218–240 (2011).
21. Hönig, S. F. *et al.* Dust in the polar region as a major contributor to the infrared emission of active galactic nuclei. *Astrophys. J.* **771,** 87 (2013).
22. Schnülle, K. *et al.* Dust physics in the nucleus of NGC 4151. *Astron. Astrophys.* **557,** L13 (2013).
23. Müller-Sánchez, F. *et al.* Outflows from active galactic nuclei: kinematics of the narrowline and coronal-line regions in Seyfert galaxies. *Astrophys. J.* **739,** 69 (2011).
24. Das, V. *et al.* Mapping the kinematics of the narrow-line region in the Seyfert galaxy NGC 4151. *Astron. J.* **130,** 945–956 (2005).
25. Martel, A. R. & New, H. Spectropolarimetry of NGC 4151: The broad-line region-host connection. *Astrophys. J.* **508,** 657–663 (1998).
26. Mundell, C. G. The nuclear regions of the Seyfert galaxy NGC 4151: parsec-scale H I absorption and a remarkable radio jet. *Astrophys. J.* **583,** 192–204 (2003).
27. Mould, J. R. *et al.* The Hubble Space Telescope Key Project on the Extragalactic Distance Scale. XXVIII. Combining the constraints on the Hubble constant. *Astrophys. J.* **529,** 786–794 (2000).
28. Hicks, E. K. S. & Malkan, M. A. Circumnuclear gas in Seyfert 1 galaxies: morphology, kinematics, and direct measurement of black hole masses. *Astrophys. J. Suppl. Ser.* **174,** 31–73 (2008).
29. Grier, C. J. *et al.* Stellar velocity dispersion measurement in high-luminosity quasar hosts and implications for the AGN black hole mass scale. *Astrophys. J.* **773,** 90 (2013).
30. Pancoast, A. *et al.* The Lick AGN Monitoring Project 2011: dynamical modeling of the broad-line region in Mrk 50. *Astrophys. J.* **754,** 49 (2012).

Acknowledgements We thank R. Wojtak for discussions on peculiar velocities near the Virgo cluster. S.F.H. acknowledges support from the Marie Curie International Incoming Fellowship within the Seventh European Community Framework Programme (PIIF-GA-2013-623804). The Dark Cosmology Centre is funded by the Danish National Research Foundation. This research has made use of the NASA/IPAC Extragalactic Database (NED) which is operated by JPL, Caltech, under contract with NASA.

Author Contributions S.F.H. and D.W. had the idea for the project. S.F.H. collected the data, developed the model and wrote the paper. D.W. assisted in interpreting the results and helped to write the manuscript. M.K. contributed the interferometry data and helped with the data analysis. J.H. contributed to the modelling and interpretation. All authors engaged in discussion and provided comments on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.F.H. (s.hoenig@soton.ac.uk).

LETTER

doi:10.1038/nature13956

An impenetrable barrier to ultrarelativistic electrons in the Van Allen radiation belts

D. N. Baker[1], A. N. Jaynes[1], V. C. Hoxie[1], R. M. Thorne[2], J. C. Foster[3], X. Li[1], J. F. Fennell[4], J. R. Wygant[5], S. G. Kanekal[6], P. J. Erickson[3], W. Kurth[7], W. Li[2], Q. Ma[2], Q. Schiller[1], L. Blum[1], D. M. Malaspina[1], A. Gerrard[8] & L. J. Lanzerotti[8]

Early observations[1,2] indicated that the Earth's Van Allen radiation belts could be separated into an inner zone dominated by high-energy protons and an outer zone dominated by high-energy electrons. Subsequent studies[3,4] showed that electrons of moderate energy (less than about one megaelectronvolt) often populate both zones, with a deep 'slot' region largely devoid of particles between them. There is a region of dense cold plasma around the Earth known as the plasmasphere, the outer boundary of which is called the plasmapause. The two-belt radiation structure was explained as arising from strong electron interactions with plasmaspheric hiss just inside the plasmapause boundary[5], with the inner edge of the outer radiation zone corresponding to the minimum plasmapause location[6]. Recent observations have revealed unexpected radiation belt morphology[7,8], especially at ultrarelativistic kinetic energies[9,10] (more than five megaelectronvolts). Here we analyse an extended data set that reveals an exceedingly sharp inner boundary for the ultrarelativistic electrons. Additional, concurrently measured data[11] reveal that this barrier to inward electron radial transport does not arise because of a physical boundary within the Earth's intrinsic magnetic field, and that inward radial diffusion is unlikely to be inhibited by scattering by electromagnetic transmitter wave fields. Rather, we suggest that exceptionally slow natural inward radial diffusion combined with weak, but persistent, wave–particle pitch angle scattering deep inside the Earth's plasmasphere can combine to create an almost impenetrable barrier through which the most energetic Van Allen belt electrons cannot migrate.

Figure 1 shows that over the first 20 months (1 September 2012 to 1 May 2014) of NASA's Van Allen Probes[11] mission lifetime, highly relativistic and ultrarelativistic electrons were present in substantial (but highly variable) numbers at a distance corresponding to a McIlwain $L \gtrsim 3$. (L is the distance in Earth radii for a magnetic field line to cross the magnetic equatorial plane in a static magnetic field model.) However, such electrons were not discernibly present at lower L values. In fact, earlier observations made by instruments on the Combined Release and Radiations Effects Satellite[12] and by the Solar, Anomalous, and Magnetospheric Particle Explorer[13–15] mission suggested that the slot region and the inner zone can be filled with electrons many megaelectronvolts in energy only following the most extreme solar wind driving conditions.

Figure 1 shows that none of the solar wind driving events (Fig. 1f, g) during the Van Allen Probes operational era transported electrons (~2 to ~10 MeV) into the region with $L < 2.8$. Also, only very occasionally did the measured plasmapause boundary ever get forced inwards as close to the Earth as $L \approx 3$ (Fig. 1c). For most of the past two years, the plasmapause has been situated beyond $L \approx 4$. The operational period of the Van Allen Probes missions has been relatively quiet geomagnetically, and the plasmasphere region often extended outwards to $L \approx 5$ or farther. Thus, contrary to prior expectations, the inner edge of the highly relativistic electron population measured by the Relativistic Electron–Proton Telescope (REPT) on board each of the two probe spacecraft (Fig. 1c) was rarely collocated with the plasmapause. Instead, an almost complete lack of very high-energy electrons (in a region of slot morphology) was seen only (but persistently) for $L < 2.8$ (Fig. 1a–e).

Figure 2 shows that the inner boundary of ultrarelativistic (~7.2 MeV) electron trapping is extremely sharp and stable for many months. Even when external solar wind driver events cause erosion of some part of the higher-L population, as on 2 September 2012 (Fig. 2a) or 1 March 2013 (Fig. 2b), the ultrarelativistic electrons remained persistently high in intensity for $L > 2.8$ and showed no measureable flux for $L < 2.8$. Furthermore, as shown in Fig. 2c, when a solar wind shock wave impinged on the magnetosphere and drove the 7.2 MeV electrons inwards in a step-like fashion on 1 October 2013, these extremely high-energy electrons moved inwards only in such a way as to again have their inner boundary at $L \approx 2.8$.

As shown in Fig. 3a–c, a sampling at 31 d intervals of radial profiles of particle fluxes over much of the Van Allen Probes lifetime shows no instance of any highly relativistic electrons migrating inwards of $L = 2.8$. The data also reveal that the boundary tends to become sharper (that is, have a steeper gradient of flux versus L) at higher electron energies. Furthermore, surveys of concurrently measured plasmaspheric hiss occurrence (Fig. 3d) from the Electric and Magnetic Field Instrument Suite and Integrated Science (EMFISIS) instrument[16] on board the probe spacecraft show no sharp boundary or radial gradient change at $L = 2.8$ for these electromagnetic waves. Thus, the presence of such a clear, persistent and seemingly impenetrable barrier to inward transport of ultrarelativistic electrons at this very specific location presents a substantial puzzle.

Multi-megaelectronvolt electron losses result from processes that scatter a particle's pitch angle into the atmospheric loss cone. For example, Earth's magnetic field exhibits the South Atlantic Anomaly (SAA), a region of weaker magnetic field strengths in a low-altitude region east of South America[17]. The effects of the expanded atmospheric loss cone associated with the SAA are centred near $L \approx 1.5$, with smaller magnetic perturbations extending outwards to $L > 3$. Therefore, precipitation of energetic electrons into the SAA would not be expected to produce the sharp boundary in trapped electrons observed at $L = 2.8$. Space weather monitoring sensor data from the Van Allen Probes confirm the presence of this sharp boundary (Extended Data Fig. 1). Data from a small, low-altitude spacecraft, the Colorado Student Space Weather Experiment (Extended Data Fig. 2), further confirm the view that an inward boundary exists at $L \approx 2.8$ for $E > 3.8$ MeV electrons that is quite separate from concurrently identified SAA effects.

Another possible reason for the region of trapped electrons to have a sharp inward boundary could relate to the precipitation of energetic electrons induced by ground-based radio transmitters. Early work[18,19] supported the view that powerful, very low-frequency (VLF) radio transmitters at fixed locations on Earth's surface could cause substantial loss

[1]Laboratory for Atmospheric and Space Physics, University of Colorado, Boulder, Colorado 80303, USA. [2]Department of Atmospheric and Oceanic Sciences, University of California, Los Angeles, California 90095, USA. [3]Massachusetts Institute of Technology, Haystack Observatory, Westford, Massachusetts 01886, USA. [4]Aerospace Corporation Space Sciences Lab, Los Angeles, California 90009, USA. [5]School of Physics and Astronomy, University of Minnesota, Minneapolis, Minnesota 55455, USA. [6]NASA Goddard Space Flight Center, Greenbelt, Maryland 20771, USA. [7]Department of Physics, University of Iowa, Iowa City, Iowa 52242, USA. [8]Center for Solar-Terrestrial Research, New Jersey Institute of Technology, Newark, New Jersey 07102, USA.

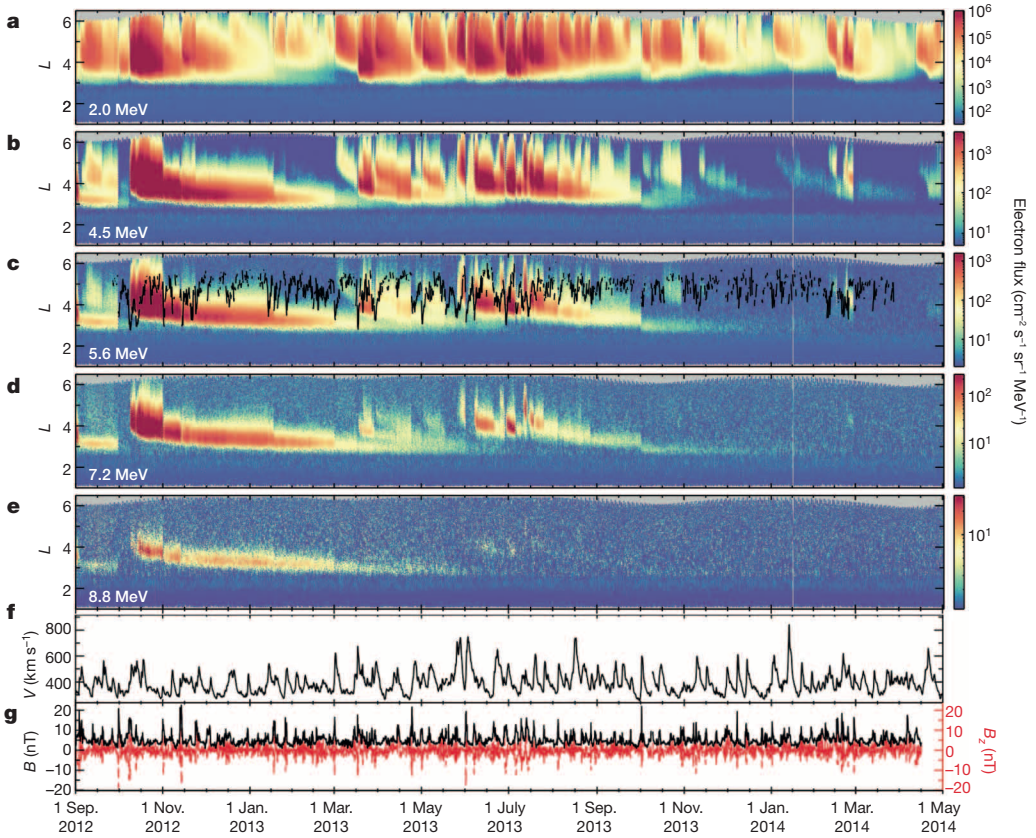


Figure 1 | A 20-month overview of electron fluxes within the Earth's Van Allen radiation belts. **a**–**e**, REPT-A instrument data[26] from initial instrument turn-on (1 September 2012) until 1 May 2014. Each panel corresponds to a different electron energy, as shown, and is plotted as a representation of electron differential-spin-averaged fluxes colour-coded as indicated. The data are shown as functions of L on the vertical axis and time along the horizontal axis. Each panel shows electron measurements from $L = 1.0$ to ~6.5 as covered by the highly elliptical Van Allen Probes[11] orbits. By comparison with similar plots of data from the same instruments in earlier studies[7], there is a smaller flux of energetic electrons for $L < 2$. This reflects a substantially improved REPT processing algorithm to remove background due to very intense inner-zone proton fluxes in all the electron channels[9]. **f**, **g**, The concurrently measured solar wind speed (**f**) and the interplanetary magnetic field (magnitude (B), black; north–south component (B_z), red) (**g**) upstream of the Earth. The broken black trace plotted over the REPT data in **c** shows the measured location of the plasmapause. This plasmapause location is derived from spacecraft potential measurements[11], which can be used as a proxy for local plasma density. We note that in **a**–**e** the highly energetic electrons measured by REPT sensors throughout the mission never seem to extend inwards of $L \approx 2.8$. This forms a particularly clear and sharp boundary for the ultrarelativistic electrons as shown in **c**–**e**.

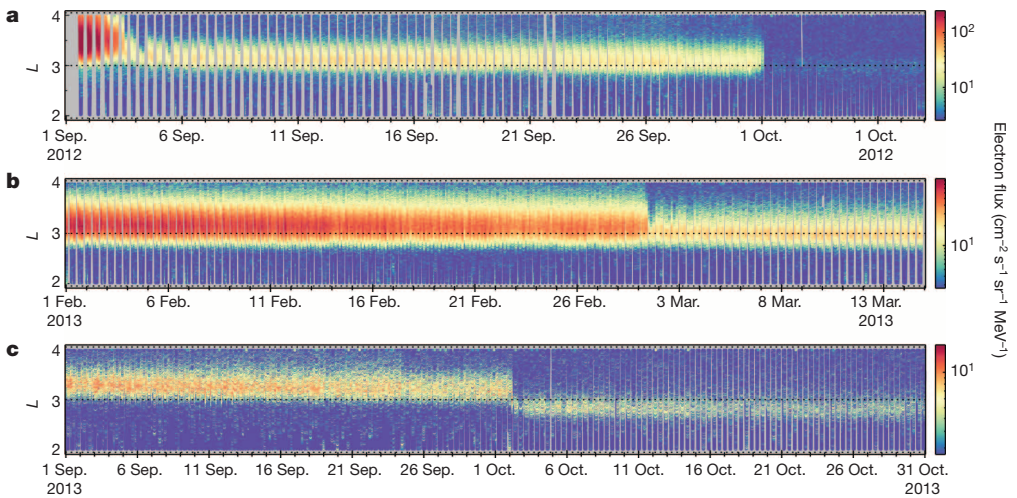


Figure 2 | Electron colour-coded data showing the sharp inner edge observed for ultrarelativistic electrons. These plots are similar to the format of Fig. 1a–e, but including the combined data from REPT-A and REPT-B[26]. The focus in this figure is upon 7.2 MeV (6.7–7.7 MeV) electrons (Fig. 1d). **a**, Sharp inner edge as seen in the L–time format for 1 September 2012 to 30 September 2013. This relativistic electron 'storage ring' disappeared suddenly on 30 September 2012[7], but before then the 7.2 MeV electrons never migrated inwards of $L = 2.8$. **b**, Similar to **a**, but showing REPT data from 1 February 2013 to 15 March 2015. Throughout this period, the inner edge of the region of 7.2 MeV electrons did not deviate from $L = 2.8$. **c**, Similar to **a**, but for the period from 1 September to 31 October 2013. From 1 September to 1 October, the 7.2 MeV electron fluxes slowly moved inwards towards $L = 3.0$. On 1 October, an abrupt solar wind transient (shock wave) drove the storage ring inwards to a new equilibrium position with the inner edge once again at $L \approx 2.8$.

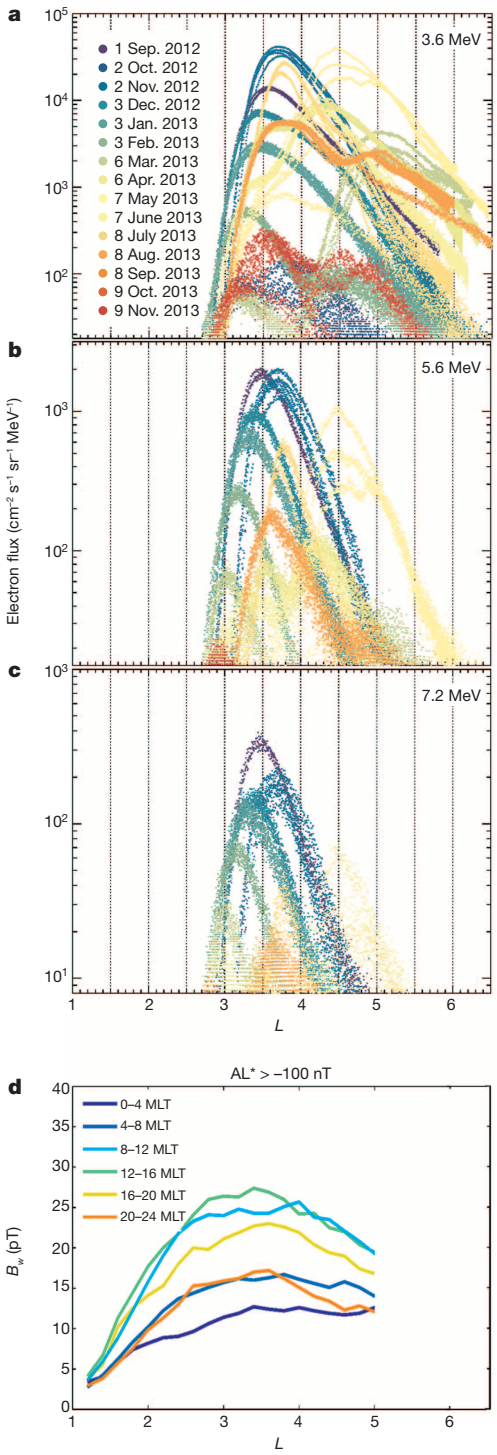


Figure 3 | Electron flux radial profiles for selected outer Van Allen zone passages. Each colour-coded profile shows the differential directional flux measured by the REPT-A instrument during a passage of a Van Allen Probe spacecraft through the magnetosphere. Passes are chosen at 31 d intervals to sample the radiation belt properties under a wide variety of conditions throughout the mission lifetime. **a**, Sampled passes for the 3.6 MeV (3.2–4.0 MeV) REPT energy channel. **b**, Similar to **a**, but for the 5.6 MeV (5.0–6.2 MeV) REPT energy channel. **c**, Similar to **a**, but for the 7.2 MeV (6.7–7.7 MeV) energy channel. In all cases, the high-energy electron profiles never extend significantly inwards of $L = 2.8$. Also, the spatial flux gradients at the inner edge of the outer Van Allen belt tend to become steeper as energy increases. **d**, Statistical survey of plasmaspheric hiss occurrence frequency for the period of the Van Allen Probes mission (based on EMFISIS data (ref. 16)). No sharp boundary is seen at $L = 2.8$ (or elsewhere) in the radial occurrence distributions of the hiss. AL* is the minimum of the auroral electrojet index within the previous 3 h time period, derived from geomagnetic variations in ground magnetometer chain data. B_w is the magnetic field intensity of the hiss wave, and MLT is magnetic local time.

of otherwise stably trapped inner-zone electrons in the 100–500 keV energy range. Although some VLF transmitter power can leak into the inner magnetosphere, previous detailed analyses have shown[19] that VLF wave interactions (due to Doppler-shifted cyclotron resonances) from ground transmitters would only be substantial for electrons with energies <0.5 MeV and would be expected to be important only at mid-latitude locations where $L < 2$. We conclude that such man-made signals do not have a substantial effect on the ultrarelativistic particles (which should instead be much more efficiently scattered by lower-frequency hiss at $L \approx 2.8$ (ref. 20)). Thus, we do not believe that the sharp boundary of electron trapping at $L = 2.8$ described above for ultrarelativistic electrons is related to VLF radio transmitters or to special features in the geomagnetic field. Moreover, REPT pitch angle plots (Extended Data Fig. 3) show no evidence of any sharp or discontinuous features in the angular distributions that might be associated with geophysical boundaries or localized sources of wave interactions.

As further shown in Extended Data Fig. 4, REPT-measured electron fluxes always peak at 90° pitch angles as the spacecraft remain near the magnetic equatorial plane (the spacecraft orbits have a 10° inclination). Therefore, strong scattering and isotropization is not evident in the data. The scattering lifetimes of highly relativistic electrons inside the plasmapause are very long (Fig. 4), and so we do not believe that the observed sharp inner edge is due to some anomalous fast loss process. Thus, we conclude that the persistent inner edge represents a remarkably stable boundary where weak scattering losses are dominant over even weaker radial diffusion transport (Fig. 4). The fact that the inner edge can move abruptly during solar disturbances (Fig. 2) suggests that its location can be changed either by a pronounced increase in the radial diffusion rate or by further local acceleration.

The inner edge of the relativistic electron population is a remarkable feature at all geographic longitudes (Extended Data Fig. 5). It has not

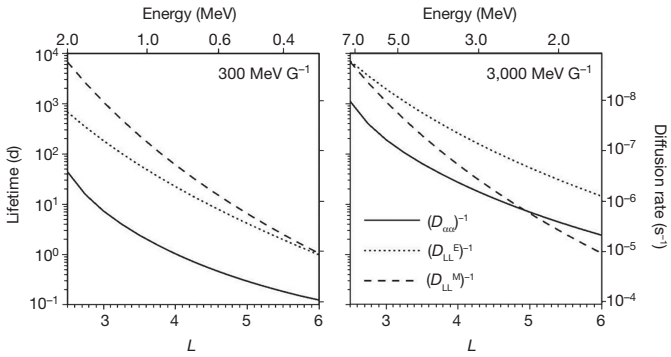


Figure 4 | A comparison between the timescales for scattering loss and inward radial diffusion. The lifetime $(D_{\alpha\alpha})^{-1}$ of energetic electrons due to pitch angle scattering by plasmaspheric hiss under weak-to-moderate geomagnetic activity (Fig. 3d), plotted (solid lines) as a function of L for two representative magnetic moments, 300 and 3,000 MeV G^{-1}. Corresponding electron energies are shown along the top axis. The rates of inward radial diffusion and the corresponding transport timescales due to ultralow-frequency fluctuating electric $(D_{LL}^{E})^{-1}$ and magnetic $(D_{LL}^{M})^{-1}$ fields, based on the parameterization of ref. 27, are shown as dotted and dashed lines, respectively. For the ultrarelativistic population (3,000 MeV G^{-1}), the timescale for radial transport inside $L \approx 3$ exceeds 1,000 d, whereas the scattering loss time is comparable to 100 d. Consequently, any ultrarelativistic electrons that are injected near $L \approx 3$ will remain in place subject only to very slow loss to the atmosphere, thus accounting for the remarkable stability of the inner edge of the ultrarelativistic rings. The lower-energy electron population (<1 MeV) near $L \approx 3$ should decay much more rapidly, on a timescale of <10 d. This is also consistent with recent observations made using the Van Allen Probes[22].

previously been discussed in the literature because we have never previously had such accurate measurements at high energies. However, this inner edge being at such low L, well inside the plasmapause, seems to require that electron acceleration occur just outside this location. The radial transport of such electrons from the heart of the outer zone to $L \approx 2.8$ is usually very slow[5] (on the timescale of years). Thus, the electrons would be significantly depleted (by several orders of magnitude) by wave scattering during inward transport from the nominal plasmapause location at around four to five Earth radii (Fig. 4). Unless there occurs a prompt interplanetary-shock-induced acceleration like the March 1991 event[21], we would favour a local wave acceleration process that occurs just outside the plasmapause[8,22] when the plasmapause is pushed into the lower-L region.

At present, we contend that the plasmapause location has a role in the formation of the inner edge, and that this requires a strong solar wind event to cause both the plasmasphere to erode through convection and the plasmapause to move to lower L. But once the ultrarelativistic electron population is formed at such low L, it will stay in place subject only to very slow decay on a timescale of 100 d (refs 23, 24). These results therefore demonstrate that extraordinarily strong spatial gradients can be maintained for quite long times in ultrarelativistic electron-trapping geometries. This has potential relevance both for plasma physics and for non-terrestrial cosmic systems that magnetically confine highly energetic particles[25].

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 18 June; accepted 10 October 2014.

1. Van Allen, J. A., Ludwig, G. H., Ray, E. C. & McIlwain, C. E. Observation of high intensity radiation by satellites 1958 alpha and gamma. *Jet Propuls.* **28,** 588–592 (1958).
2. Van Allen, J. A. & Frank, L. A. Radiation around the Earth to a radial distance of 107,400 km. *Nature* **183,** 430–434 (1959).
3. Johnson, M. H. & Kierein, J. Combined Release and Radiation Effects Satellite (CRRES): spacecraft and mission. *J. Spacecr. Rockets* **29,** 556–563 (1992).
4. Blake, J. B., Baker, D. N., Turner, N., Ogilvie, K. W. & Lepping, R. P. Correlation of changes in the outer-zone relativistic electron population with upstream solar wind and magnetic field measurements. *Geophys. Res. Lett.* **24,** 927–929 (1997).
5. Lyons, L. R. & Thorne, R. M. Equilibrium structure of radiation belt electrons. *J. Geophys. Res.* **78,** 2142–2149 (1973).
6. Li, X., Baker, D. N., O'Brien, P., Xie, L. & Zong, Q. G. Correlation between the inner edge of outer radiation belt electrons and the innermost plasmapause location. *Geophys. Res. Lett.* **33,** L14107 (2006).
7. Baker, D. N. *et al.* A long-lived relativistic electron storage ring embedded in Earth's outer Van Allen belt. *Science* **340,** 186–190 (2013).
8. Thorne, R. M. *et al.* Rapid local acceleration of relativistic radiation belt electrons by magnetospheric chorus. *Nature* **504,** 411–414 (2013).
9. Baker, D. N. *et al.* The Relativistic Electron-Proton Telescope (REPT) instrument on board the Radiation Belt Storm Probes (RBSP) spacecraft: characterization of Earth's radiation belt high-energy particle populations. *Space Sci. Rev.* **179,** 337–381 (2013).
10. Spence, H. E. *et al.* Science goals and overview of the Radiation Belt Storm Probes (RBSP) Energetic Particle, Composition, and Thermal Plasma (ECT) suite on NASA's Van Allen Probes mission. *Space Sci. Rev.* **179,** 311–336 (2013).
11. Mauk, B. H. *et al.* Science objective and rationale for the Radiation Belt Storm Probes mission. *Space Sci. Rev.* **179,** 3–27 (2013).
12. Blake, J. B., Kolasinski, W. A., Fillius, R. W. & Mullen, E. G. Injection of electrons and protons with energies of tens of MeV into $L < 3$ on March 24, 1991. *Geophys. Res. Lett.* **19,** 821–824 (1992).
13. Baker, D. N. *et al.* An extreme distortion of the Van Allen belt arising from the 'Hallowe'en' solar storm in 2003. *Nature* **432,** 878–881 (2004).
14. Baker, D. N., Kanekal, S. G., Horne, R. B., Meredith, N. P. & Glauert, S. A. Low-altitude measurements of 2–6 MeV electron trapping lifetimes at $1.5 \leq L \leq 2.5$. *Geophys. Res. Lett.* **34,** L20110 (2007).
15. Zhao, H. & Li, X. Inward shift of outer radiation belt electrons as a function of *Dst* index and the influence of the solar wind on electron injections into the slot region. *J. Geophys. Res. Space Phys.* **118,** 756–764 (2013).
16. Kletzing, C. A. *et al.* The Electric and Magnetic Field Instrument Suite and Integrated Science (EMFISIS) on RBSP. *Space Sci. Rev.* **179,** 127–181 (2013).
17. Roederer, J. G. *Dynamics of Geomagnetically Trapped Radiation* (Springer, 1970).
18. Vampola, A. L. & Kuck, G. A. Induced precipitation of inner zone electrons, 1. Observations. *J. Geophys. Res.* **83,** 2543–2551 (1978).
19. Koons, H. C., Edgar, B. C. & Vampola, A. L. Precipitation of inner zone electrons by Whistler mode waves from the VLF transmitters UMS and NWC. *J. Geophys. Res.* **86,** 640–648 (1981).
20. Abel, R. W. & Thorne, R. M. Electron scattering loss in Earth's inner magnetosphere: 1. Dominant physical processes. *J. Geophys. Res.* **103,** 2385–2396 (1998).
21. Li, X. *et al.* Simulation of the prompt energization and transport of radiation particles during the March 23, 1991 SSC. *Geophys. Res. Lett.* **20,** 2423–2426 (1993).
22. Baker, D. N. *et al.* Gradual diffusion and punctuated phase space density enhancements of highly relativistic electrons: Van Allen Probes observations. *Geophys. Res. Lett.* **41,** 1351–1358 (2014).
23. Ni, B., Bortnik, J., Thorne, R. M., Ma, Q. & Chen, L. Resonant scattering and resultant pitch angle evolution of relativistic electrons by plasmaspheric hiss. *J. Geophys. Res.* **118,** 7740–7751 (2013).
24. Thorne, R. M. *et al.* Evolution and slow decay of an unusual narrow ring of relativistic electrons near $L \sim 3.2$ following the September 2012 magnetic storm. *Geophys. Res. Lett.* **40,** 3507 (2013).
25. Eastlund, B. J., Miller, B. & Michel, F. C. Emission from closed and filled magnetospheric shells and its application to the Crab pulsar. *Astrophys. J.* **483,** 857–867 (1997).
26. NASA Goddard Space Flight Center. Coordinated Data Analysis Web, http://cdaweb.gsfc.nasa.gov/istp_public/ (11 September 2014).
27. Brautigam, D. H. & Albert, J. Radial diffusion analysis of outer radiation belt electrons during the October 9, 1990 magnetic storm. *J. Geophys. Res.* **105,** 291–309 (2000).

Acknowledgements We thank the entire Van Allen Probes mission team for suggestions about this work. Data access was provided through the Johns Hopkins University/Applied Physics Lab Mission Operations Center and the Los Alamos National Laboratory Science Operations Center. This work was supported by JHU/APL contract 967399 under NASA's prime contract NAS5-01072. All Van Allen Probes data used are publicly available at http://www.rbsp-ect.lanl.gov.

Author Contributions D.N.B. developed the project, directed the data analysis and was primarily responsible for writing the paper. A.N.J., V.C.H. and S.G.K. analysed REPT data and produced related figures. R.M.T. provided theoretical guidance. J.C.F. and P.J.E. provided ground-based data for context. J.F.F. provided access to supplementary Van Allen Probes particle data. X.L., L.B. and Q.S. provided REPTile data. D.M.M. provided plasmapause location from EFW data. J.R.W. provided electric field data and W.K. provided EMFISIS data access. W.L. performed hiss data statistical analysis. Q.M. performed particle scattering and diffusion lifetime calculations. A.G. and L.J.L. provided ERM data from the Van Allen Probes mission.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.N.B. (daniel.baker@lasp.colorado.edu).

LETTER

doi:10.1038/nature13865

Metallization of vanadium dioxide driven by large phonon entropy

John D. Budai[1]*, Jiawang Hong[1]*, Michael E. Manley[1], Eliot D. Specht[1], Chen W. Li[1], Jonathan Z. Tischler[2], Douglas L. Abernathy[3], Ayman H. Said[2], Bogdan M. Leu[2], Lynn A. Boatner[1], Robert J. McQueeney[4] & Olivier Delaire[1]

Phase competition underlies many remarkable and technologically important phenomena in transition metal oxides. Vanadium dioxide (VO_2) exhibits a first-order metal–insulator transition (MIT) near room temperature, where conductivity is suppressed and the lattice changes from tetragonal to monoclinic on cooling. Ongoing attempts to explain this coupled structural and electronic transition begin with two alternative starting points: a Peierls MIT driven by instabilities in electron–lattice dynamics and a Mott MIT where strong electron–electron correlations drive charge localization[1–10]. A key missing piece of the VO_2 puzzle is the role of lattice vibrations. Moreover, a comprehensive thermodynamic treatment must integrate both entropic and energetic aspects of the transition. Here we report that the entropy driving the MIT in VO_2 is dominated by strongly anharmonic phonons rather than electronic contributions, and provide a direct determination of phonon dispersions. Our *ab initio* calculations identify softer bonding in the tetragonal phase, relative to the monoclinic phase, as the origin of the large vibrational entropy stabilizing the metallic rutile phase. They further reveal how a balance between higher entropy in the metal and orbital-driven lower energy in the insulator fully describes the thermodynamic forces controlling the MIT. Our study illustrates the critical role of anharmonic lattice dynamics in metal oxide phase competition, and provides guidance for the predictive design of new materials.

Vanadium dioxide is both an archetypical example of competing metallic and insulating phases[1–10] and a material useful in applications such as field-effect transistors and ultrafast optoelectronic switches[10–12]. In the metallic tetragonal (rutile) phase above $T_c = 340$ K, vanadium atoms form straight, equally spaced *c*-axis chains, whereas in the insulating monoclinic (M1) phase, they dimerize to form zigzag chains with unequal spacings (Fig. 1c). Concurrently, an insulating gap opens in the electronic structure. Numerous studies indicate that changes in either electron–lattice (Peierls) or electron–electron (Mott) correlations can impact physical properties, particularly when doping or under external strains. However, our basic understanding of the MIT in VO_2 lacks an accurate description of lattice dynamics because the incoherent vanadium neutron scattering cross-section hinders single-crystal inelastic neutron scattering (INS). Thus, attempts to elucidate VO_2 lattice dynamics have relied on indirect approaches such as Raman scattering[13], thermal diffuse scattering[14] (TDS), Debye–Waller measurements[15], acoustic waves[16] and INS studies of related materials[17]. These results identified large vibrational amplitudes in the metallic phase[15] and suggested a soft-mode phase transition associated with a particular lattice periodicity[14,18] (the tetragonal R-wavevector in reciprocal space). However, they provided only a rudimentary picture of lattice dynamics, preventing a definitive assessment of thermodynamic forces driving the MIT[6,9,16,18,19]. Here we report X-ray and neutron scattering measurements and *ab initio* molecular dynamics (AIMD) calculations that provide a full thermodynamic accounting.

We first examine lattice vibrations in VO_2 using measurements of the temperature-dependent phonon density of states (PDOS), made using the ARCS neutron spectrometer at the Spallation Neutron Source at Oak Ridge National Laboratory. The PDOS spectra shift gradually above and below the MIT, but exhibit a discontinuity in shape across T_c (Fig. 1a). The rutile PDOS at temperatures above T_c are considerably softer (lower in energy) than the M1-phase PDOS. In addition, a low-energy peak is evident in the momentum-averaged rutile dynamical structure factor $S_{Qavg}(E)$ around 14 meV at high temperatures (1,200 K; Fig. 1d). This peak gradually softens by a few millielectronvolts on cooling towards T_c. From our inelastic X-ray scattering (IXS) dispersion measurements and calculations (discussed below), this peak is identified with transverse acoustic branches, which are flat (constant in energy) over large portions of the Brillouin zone. On cooling to near T_c, the energy of this peak remains above ~12 meV and then disappears as VO_2 transforms to the M1 phase in a first-order transition (Fig. 1d). Importantly, the changes in the PDOS provide the phonon contribution to the transition entropy. Our experimental vibrational entropy change, corrected for neutron weighting, is $\Delta S_{ph}(\text{rutile–M1}) = 0.34 \pm 0.03\ k_B\ \text{atom}^{-1}$ (k_B, Boltzmann's constant). Latent heat measurements yield $\Delta H = T_c \Delta S = 14.7$ meV atom^{-1}, or $\Delta S = 0.5 \pm 0.05\ k_B\ \text{atom}^{-1}$, for the total entropy change[4,9]. Thus, phonons account for ~2/3 of the total increase in entropy at the MIT, dominating the entropy stabilizing the metal. This result is consistent with early speculation[6], rules out other estimates[5,16,20] and, most importantly, provides a quantitative benchmark for theory.

To explain this experimental result, we calculated temperature-dependent phonon dispersions with density functional theory (DFT). We used the Perdew–Burke–Ernzerhof (PBE) exchange-correlation functional, and the modification PBE+*U* that includes an electronic Hubbard repulsion *U* on vanadium sites as suggested by previous calculations[19,21]. Additional calculations with the Heyd–Scuseria–Ernzerhof (HSE) hybrid functional treated electronic correlations at the level of PBE ($U = 0$), but with a more exact exchange contribution. Simulations including strong anharmonicity in the rutile phase were performed using AIMD (PBE) and the temperature-dependent effective potential method[22]. The M1 phase was found to be largely harmonic, and its phonons were calculated in this approximation. The calculated PDOSs for the M1 ($T = 0$ K) and rutile ($T = 425$ K) phases are shown for comparison in Fig. 1b. Good overall agreement is found between the calculated and measured PDOSs with regard to both the shapes and the energies of the main features, including the rutile transverse acoustic peak at ~12 meV (near T_c) and the relative softness of the metallic phase.

Using *ab initio* calculations, we computed the contributions of phonons and electrons to the entropies of both phases. The computed vibrational entropy change at the transition yields 0.31 k_B atom^{-1} for AIMD (PBE, $U = 0$), showing remarkable agreement with experiment. The electronic entropy change was obtained from the electronic density of states (EDOS), given a sufficient bandgap in the M1 phase for its electronic

[1]Materials Science and Technology Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA. [2]Advanced Photon Source, Argonne National Laboratory, Argonne, Illinois 60439, USA. [3]Quantum Condensed Matter Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA. [4]Neutron Sciences Directorate, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA.

*These authors contributed equally to this work.

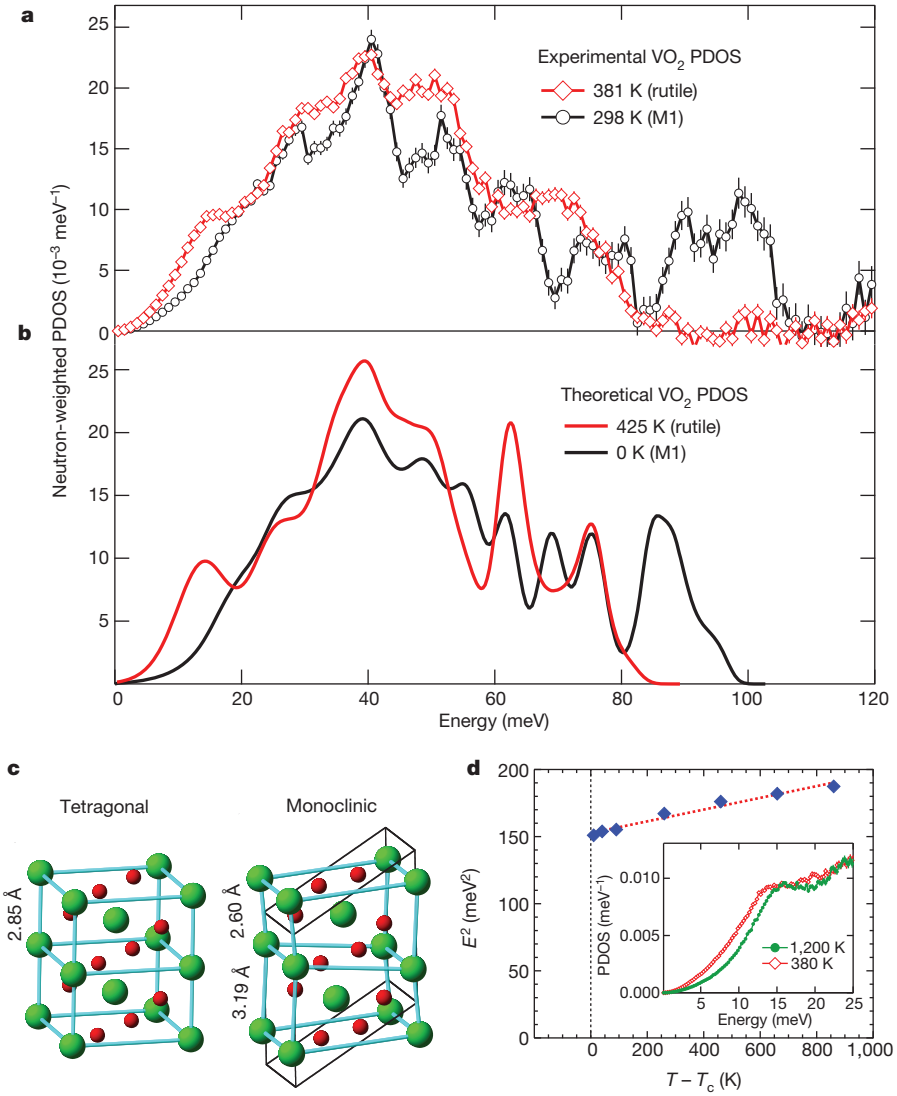
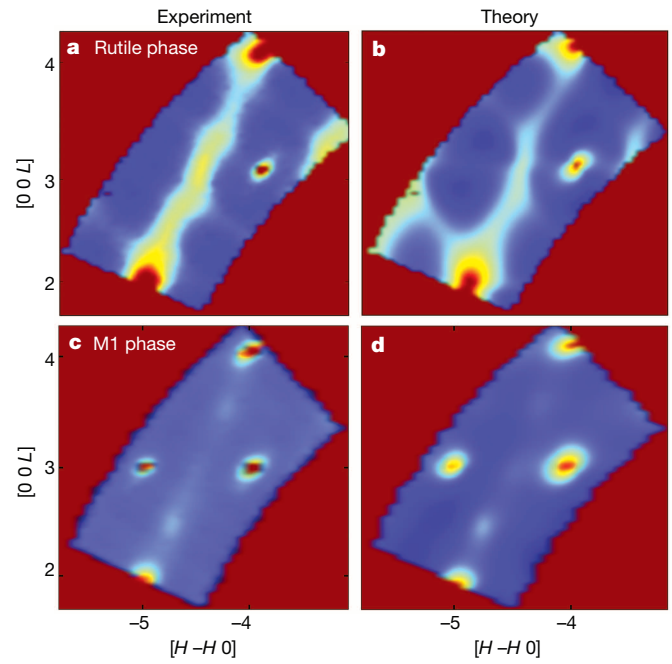


Figure 1 | PDOS spectra of rutile and M1 phases in VO_2 showing phonon stiffening in the M1 phase. **a**, Neutron-weighted PDOSs measured with INS in the rutile phase at 381 K and in the M1 phase at 298 K (error bars denote standard deviation from counting statistics). **b**, PDOSs computed using the temperature-dependent effective potential method for the rutile phase at 425 K and from DFT for the M1 phase at 0 K. **c**, VO_2 atomic structures in the tetragonal (rutile) and monoclinic (M1) phases (V, green; O, red). **d**, Temperature dependence of low-energy rutile transverse acoustic peak plotted as squared energy ($\hbar^2\omega^2$)versus temperature offset from the transition temperature T_c, showing gradual softening and abrupt disappearance at T_c. Inset shows PDOSs at 380 and 1,200 K.



entropy to be negligible (in agreement with our DFT calculations and the known insulating nature of the M1 phase). The rutile EDOS gave S_{el}(rutile−M1) = 0.09 k_B atom^{-1} for both PBE and PBE+U, in agreement with transport estimates[9]. Thus, our best estimate of the total entropy change across the transition is $\Delta S_{tot} = \Delta S_{ph}(exp) + \Delta S_{el}(calc) =$ 0.43 k_B atom^{-1}. This result accounts for most of the calorimetry total of 0.5 k_B atom^{-1} (refs 4, 9), and shows that other possible contributions (from, for example, magnetism) are small.

To identify the origin of large vibrational entropy for particular wavevectors, $\boldsymbol{q}$, in reciprocal space, we performed X-ray TDS measurements[14] from VO_2 single crystals at Advanced Photon Source (APS) beamline 33-BM. Figure 2a, c shows the measured diffuse intensity in a rutile {110} reciprocal-space slice, and Fig. 2b, d shows the corresponding TDS patterns computed from the *ab initio* phonon dispersions. The calculations agree well with the measured TDS $\boldsymbol{q}$ profiles, including subtle intensity variations, demonstrating that the TDS streak in Fig. 2 is due to low energy phonons. An observed linear increase in rutile TDS

Figure 2 | Comparison of experimental and calculated X-ray TDS. **a**, **b**, Rutile {110} slice in reciprocal space defined by $H + K = 0$, measured at a temperature of 358 K (**a**) and calculated at 425 K (**b**). **c**, **d**, Same reciprocal-space slice as in **a** and **b**, but in the M1 phase, measured at 338 K (**c**) and calculated at 320 K (**d**). The strong plane of diffuse intensity abruptly disappears below the MIT temperature. Colours indicate scattered intensity based on logarithmic rainbow scales in arbitrary units, with red as the highest intensity. H, K and L refer to rutile coordinates.

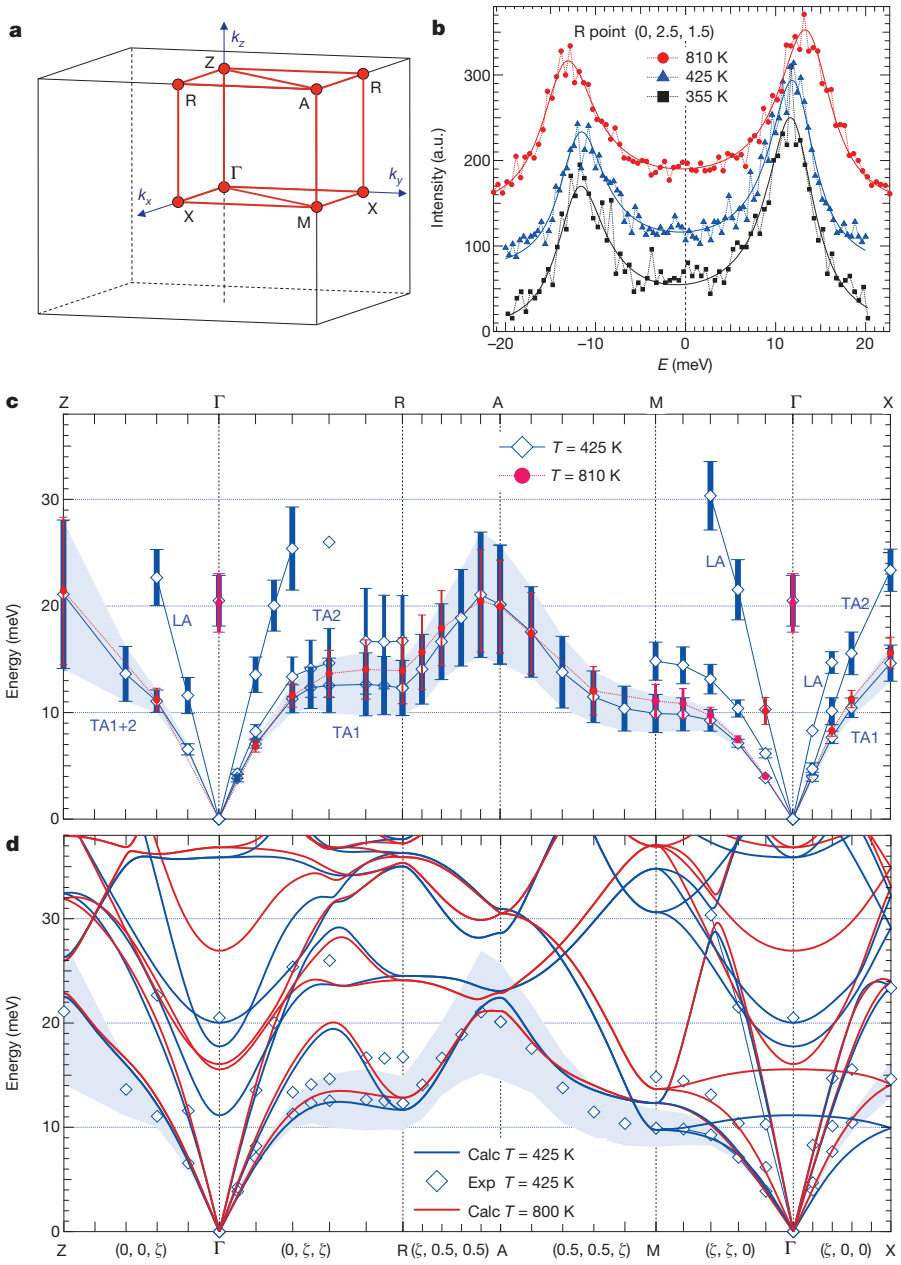


Figure 3 | Phonon dispersions in rutile VO_2. **a**, Reciprocal-lattice labels for tetragonal VO_2. **b**, Constant-$\boldsymbol{q}$ IXS scans at the rutile R point showing the broad energy widths are well fitted using damped oscillator line shapes. a.u., arbitrary units. **c**, Experimental IXS low-energy phonon dispersions at 425 and 800 K showing flat transverse acoustic branches that stiffen on heating. Large energy widths are indicated by vertical lines and shading. The various transverse acoustic (TA) and longitudinal acoustic (LA) modes are labelled accordingly. By symmetry, TA1 and TA2 are equivalent along Γ–Z and are labelled as TA1 + 2. **d**, Lines show phonon dispersions for rutile VO_2 calculated from *ab initio* anharmonic lattice dynamics at $T = 425$ K (blue) and 800 K (red). Diamonds and shading are experimental phonon energies and widths measured with IXS at 425 K.

intensity with temperature further confirms that it is due to phonons as opposed to static displacements. The abrupt disappearance of the diffuse signal in the M1 phase reflects the phonon stiffening and disappearance of the low-energy PDOS transverse acoustic peak. The three-dimensional (3D) rutile TDS profiles reveal that soft phonons are concentrated in rutile {111} sheets with Miller indices H, K and L satisfying the relation $H + K + L = 2n$ (n a non-zero integer) (Supplementary Video 1). This observation is compatible with measurements in a related compound[23], $V_{0.9}Nb_{0.1}O_2$, but is inconsistent with diffuse VO_2 lobes previously inferred[14]. The {111} TDS sheets pass through R and M points in the tetragonal Brillouin zone (Fig. 3a), and have been proposed to arise from correlated linear displacements of chains of vanadium atoms in the real-space [111] direction[23]. However, energy-integrated TDS measurements alone cannot determine the nature of the displacements. We now trace the {111} TDS to particular low-energy transverse acoustic phonons.

Energy- and momentum-resolved IXS measurements at the APS HERIX beamline revealed the origin of individual low-energy modes, and, in addition, revealed strong damping of rutile phonons. As shown in Fig. 3b, energy scans at constant $\boldsymbol{q}$ reveal unusually broad phonon peaks (short lifetimes) with asymmetric shapes matching damped oscillator line shapes. Figure 3c shows experimental rutile dispersions where the phonon scattering rates, $2\gamma_j(\boldsymbol{q})$, are indicated as vertical bars. Vanadium atom motions dominate the lowest-energy branches. Flat transverse acoustic sections along Γ–R and Γ–M ($E \approx 12 \pm 3$ meV) are responsible for both the low-energy PDOS peak and the {111} TDS sheets described above. We find that all low-energy transverse acoustic and longitudinal acoustic phonons are strongly damped, particularly along Γ–R–A–M. Furthermore, unusual stiffening of the low-energy transverse acoustic branches is observed on heating from 425 to 800 K. This stiffening is strongest along Γ–R and Γ–M, suggesting that the same anharmonic mechanism is responsible for both stiffening and large damping. The phonon scattering rates are large, corresponding to low quality factors $Q = E/2\gamma \approx 3$. Notably, the large phonon linewidths are largely temperature independent. This unusual observation suggests an offsetting competition between electron–phonon interactions (increasing Peierls coupling and widths on cooling) and conventional anharmonicity (increasing phonon–phonon interactions on heating). We note that the

R-point transverse acoustic phonon energy does not trend to zero during cooling close to the MIT; rather, the energy remains above ~12 meV (Fig. 1d). This observation disproves the proposal, suggested by symmetry, of a conventional soft-mode transition at this wavevector[14,18], and shows that harmonic DFT attempts focused on this lattice instability do not capture the pervasive nature of anharmonic lattice dynamics[19].

We note that rutile VO_2 phonon dispersions show a strong similarity to the dispersions in rutile TiO_2 (ref. 24), although the VO_2 PDOS is shifted down in energy. This softness explains the anomalously large thermal amplitudes observed in VO_2: the root mean squared vibrational displacements of vanadium atoms above the MIT is 0.18 Å (ref. 15), nearly as large as the 0.21 Å static displacements when transforming from rutile to M1, and much larger than vibrational amplitudes in the VO_2 M1 phase and the TiO_2 rutile phase[15]. Our calculated amplitudes in rutile and M_1 VO_2 are in agreement with the experimental trends (Methods and Extended Data Table 1). Because the crystal structure is the same and the atomic masses are similar, this phonon softening in VO_2 compared with TiO_2 could be ascribed to the difference in bonding between the two materials. Titanium dioxide is a wide-gap insulator (>3 eV), but in VO_2 the extra *d* electron of vanadium produces a metallic state, partly filling the bands derived from the vanadium $3d\ t_{2g}$ orbitals. Metallic behaviour screens the phonons, lowering their frequencies[6].

Using calculations to understand the origin of soft phonons and strong anharmonicity (Fig. 3b, c), we first observe that rutile VO_2 dispersions calculated in the harmonic approximation produced unstable phonons across a wide region of the Brillouin zone (both PBE and PBE+*U*), consistent with prior reports[19]. In contrast, our AIMD dispersion calculations for rutile VO_2 (Fig. 3d), including renormalization by anharmonicity, show good agreement with experiment, without any fitting parameters. Although the rutile phase is dynamically unstable at 0 K, our results establish that anharmonic bonding stabilizes the rutile lattice at temperatures above the MIT, producing stable phonons throughout the Brillouin zone. Consistent with IXS (Fig. 3c) and TDS (Fig. 2), calculated phonons are soft for wavevectors in reciprocal-space {111} planes that include the tetragonal R and M points. We calculate that increasing the temperature to 800 K stiffens the phonons at a rate of 3.3 μeV K^{-1} at the R point, in agreement with the measured 4 μeV K^{-1}.

Anharmonicity and softness in the rutile phase are revealed with DFT calculations of frozen-phonon potential energy curves. Figure 4a shows the energy profile (per atom) for the particular rutile R-point transverse acoustic modes (0.5TA1 + 0.5TA2) associated with the static atomic displacements that occur during the MIT, and Fig. 4b shows the profile for the related optic mode at the corresponding wavevector (Γ point) in the M1 phase. The quartic-like potential for all low-energy rutile modes strongly departs from parabolic (harmonic) behaviour, manifesting the nonlinear dependence of ionic forces on displacement amplitudes. In contrast, the M1 phase shows harmonic frozen-phonon potentials for all modes calculated. This agrees with Raman spectra being sharp in the M1 phase and broad in the rutile phase, suggesting strong metallic electron–phonon coupling[6,13]. Thermodynamically, the anharmonic potentials describe the microstructural mechanism stabilizing the high-temperature rutile structure. Phase stability is controlled by competition between minimizing enthalpy, *H*, and maximizing entropy, *S*, in the Gibbs free energy, $G = H - TS$. The soft, quartic rutile potentials significantly increase the position–momentum phase space volume, providing the large vibrational entropy that stabilizes the metallic phase.

Illustrating the effects of structural and dynamic changes, Fig. 4c, d shows how the total energy and the forces on vanadium atoms change as the lattice structure is transformed from rutile to M1. Here intermediate configurations of the schematic form $\xi \times \mathrm{M1} + (1-\xi) \times \mathrm{rutile}$ are taken as the linear interpolation of atomic positions and lattice parameters that are described by the continuous parameter ξ, with lattice structures ranging from rutile ($\xi = 0$) to M1 ($\xi = 1$). For the DFT calculation with no Hubbard repulsion ($U = 0$), there are two shallow minima with comparable energies, one in the rutile phase and one in the M1 phase. Including electronic correlations, the PBE+*U* calculation

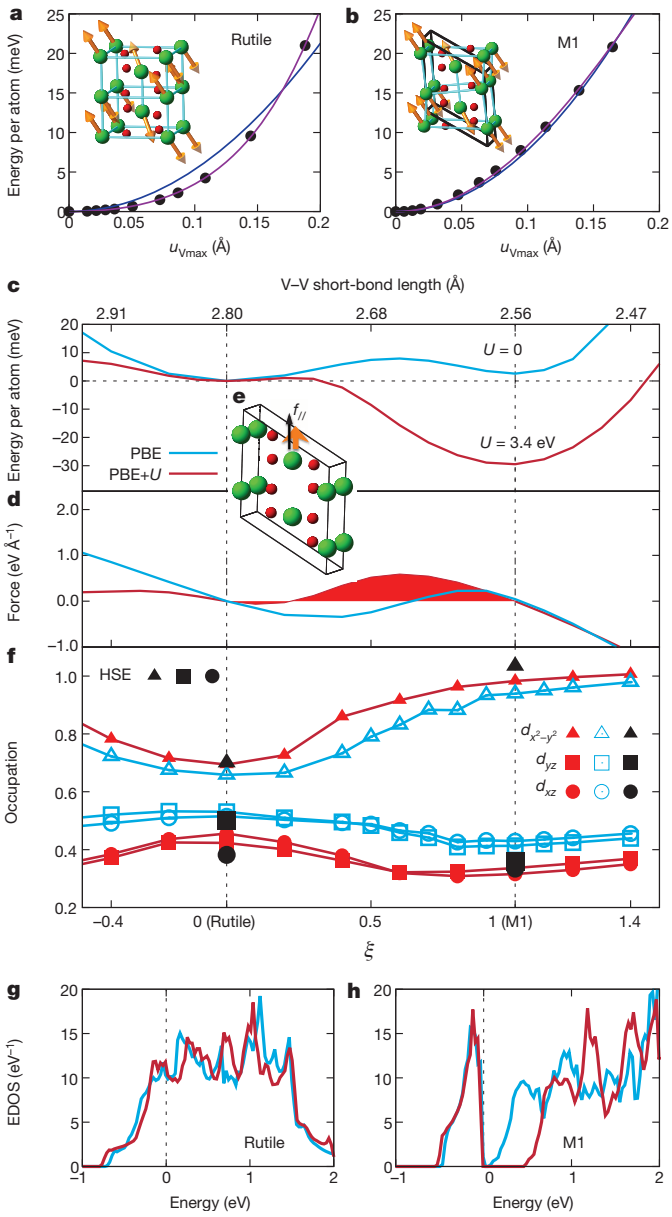


Figure 4 | Simulations of VO_2 phases. **a**, **b**, Quadratic (blue) and quartic (magenta) DFT frozen-phonon potentials for rutile R-point (**a**) and M1 Γ-point (**b**) modes with vanadium atom displacements shown inset. Calculated energy (solid circles) is shown versus maximum vanadium amplitude, u_{Vmax}. **c**, **d**, **f**, Energies (**c**), forces ($f_{//}$; **d**) and orbital occupancies (**f**), for intermediate-structures between the rutile ($\xi = 0$) and M1 ($\xi = 1$) phases. **e**, Displacements (brown arrow) along *c*-axis vanadium chain, corresponding to the forces in **d**. **g**, **h**, EDOSs in the metallic (**g**) and insulating (**h**) phases ($E_{Fermi} = 0$ meV). In **c**, **d**, **f**–**h**, red and light blue curves represent PBE+*U* and PBE results, respectively; in **f**, black symbols represent HSE calculations.

exhibits a lower minimum for the M1 phase, corresponding to the energetically favoured insulating phase at low temperature. A metastable low-temperature rutile structure with a small energy barrier (Fig. 4c) is similar to results plotted in ref. 7. For PBE+*U*, the energy difference between the rutile and M1 structures is 29 meV $atom^{-1}$, larger than experiment[4,9] (15 meV $atom^{-1}$) and a previous local-density approximation result[7]. A smaller value of $U < 3.4$ eV would produce the experimental energy difference, although with a smaller bandgap. Examining the corresponding atomic forces using PBE+*U* shows that the force on a vanadium atom along the *c*-axis chain, $f_{//}$ (schematically indicated in Fig. 4e), is initially slightly negative (restoring) near $\xi = 0$ (rutile), but becomes positive (destabilizing) for larger ξ values (shaded region Fig. 4d).

Thus, V–V dimerization is favoured by Peierls distortion, energetically stabilizing the M1 phase. Our DFT calculations with $U = 0$ show a weaker Peierls instability, with a positive force only for $\xi > 0.64$.

Motivated by suggestions that the Peierls distortion is orbitally driven[19,25–29], in Fig. 4f we show that the vanadium $d_{x^2-y^2}$ orbital occupancy increases while occupation in the other two t_{2g} orbitals decreases as the lattice transforms from rutile to M1, increasing anisotropy. We note that the Peierls instability begins once the difference in occupation between the $d_{x^2-y^2}$ and d_{yz} (or d_{xz}) orbitals exceeds ~0.4 electrons in both PBE and PBE+U calculations. Additional hybrid functional (HSE) calculations (Fig. 4f) show that both PBE+U and HSE enhance the occupation of $d_{x^2-y^2}$ orbitals relative to PBE ($U = 0$), leading to stronger orbital overlap along c-axis chains and, consequently, a stronger Peierls instability. Thus, strong Hubbard electronic correlations, which are not included in HSE, are not required to describe occupations[30]. A further theoretical test is correctly predicting the EDOS for both phases. Figure 4g, h contrasts the PBE+U metallic EDOS (non-zero at Fermi level) with the insulating M1 EDOS (open bandgap). In contrast, we find a very small M1 bandgap (<0.1 eV) in relaxed PBE ($U = 0$), whereas our HSE calculation yields a gap of 1.1 eV. Thus, both PBE+U and HSE increase the magnitude of the M1 gap compared with PBE. This larger gap is also understood from the higher occupation of $d_{x^2-y^2}$ orbitals in both PBE+U and HSE, favouring the orbital-assisted Peierls distortion in agreement with references [19,25–29].

Our measurements establish fundamental benchmarks to test and guide theoretical models of the MIT in VO_2. Our first-principles calculations provide a comprehensive thermodynamic description, revealing that increased occupation of vanadium $d_{x^2-y^2}$ orbitals triggers the Peierls instability, lowering the energy and opening the insulating bandgap. The MIT results from the competition between lower electronic energy in insulating the M1 phase due to the Peierls instability, and the higher entropy of the metallic rutile phase resulting from soft anharmonic phonons. Soft lattice dynamics in the rutile phase reflect the intrinsic influence of the electronic and structural instabilities driving the MIT. This understanding of the role of lattice dynamics and their relationship to electronic structure provides a critical component for developing more complete physical models of phase competition in functional transition metal oxides.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 13 June; accepted 12 September 2014.
Published online 10 November 2014.

1. Goodenough, J. B. The two components of crystallographic transition in VO_2. *J. Solid State Chem.* **3,** 490–500 (1971).
2. Whittaker, L., Patridge, C. J. & Banerjee, S. Microscopic and nanoscale perspective of the metal-insulator phase transitions of VO_2: some new twists to an old tale. *J. Phys. Chem. Lett.* **2,** 745–758 (2011).
3. Eyert, V. The metal-insulator transitions of VO_2: a band theoretical approach. *Ann. Phys. (Leipzig)* **11,** 650–704 (2002).
4. Park, J. H. *et al.* Measurement of a solid-state triple point at the metal-insulator transition in VO_2. *Nature* **500,** 431–434 (2013).
5. Zylbersztejn, A. & Mott, N. F. Metal-insulator transition in vanadium dioxide. *Phys. Rev. B* **11,** 4383–4395 (1975).
6. Hearn, C. J. Phonon softening and metal-insulator transition in VO_2. *J. Phys. C* **5,** 1317–1334 (1972).
7. Wentzcovitch, R. M., Schulz, W. W. & Allen, P. B. VO_2 - Peierls or Mott-Hubbard - a view from band theory. *Phys. Rev. Lett.* **72,** 3389–3392 (1994).
8. Rice, T. M. *et al.* Comment and reply on 'VO_2 - Peierls or Mott-Hubbard - a view from band theory'. *Phys. Rev. Lett.* **73,** 3042–3043 (1994).
9. Berglund, C. N. & Guggenheim, H. J. Electronic properties of VO_2 near the semiconductor-metal transition. *Phys. Rev.* **185,** 1022–1033 (1969).
10. Nakano, M. *et al.* Collective bulk carrier delocalization driven by electrostatic surface charge accumulation. *Nature* **487,** 459–462 (2012).
11. Aetukuri, N. B. *et al.* Control of the metal-insulator transition in vanadium dioxide by modifying orbital occupancy. *Nature Phys.* **9,** 661–666 (2013).
12. Jeong, J. *et al.* Suppression of metal-insulator transition in VO_2 by electric field-induced oxygen vacancy formation. *Science* **339,** 1402–1405 (2013).
13. Srivastava, R. & Chase, L. L. Raman spectrum of semiconducting and metallic VO_2. *Phys. Rev. Lett.* **27,** 727–730 (1971).
14. Terauchi, H. & Cohen, J. B. Diffuse X-ray-scattering due to lattice instability near the metal-semiconductor transition in VO_2. *Phys. Rev. B* **17,** 2494–2496 (1978).
15. McWhan, D. B., Marezio, M., Remeika, J. P. & Dernier, P. D. X-ray-diffraction study of metallic VO_2. *Phys. Rev. B* **10,** 490–495 (1974).
16. Maurer, D., Leue, A., Heichele, R. & Müller, V. Elastic behavior near the metal-insulator transition of VO_2. *Phys. Rev. B* **60,** 13249–13252 (1999).
17. Pynn, R., Axe, J. D. & Raccah, P. M. Structural fluctuations in NbO_2 at high temperatures. *Phys. Rev. B* **17,** 2196–2205 (1978).
18. Gervais, F. & Kress, W. Lattice-dynamics of oxides with rutile structure and instabilities at the metal-semiconductor phase-transitions of NbO_2 and VO_2. *Phys. Rev. B* **31,** 4809–4814 (1985).
19. Kim, S., Kim, K., Kang, C. J. & Min, B. I. Correlation-assisted phonon softening and the orbital-selective Peierls transition in VO_2. *Phys. Rev. B* **87,** 195106 (2013).
20. Pintchovski, F., Glaunsinger, W. S. & Navrotsky, A. Experimental study of electronic and lattice contributions to VO_2 transition. *J. Phys. Chem. Solids* **39,** 941–949 (1978).
21. Qu, B. Y., He, H. Y. & Pan, B. C. The dynamical process of the phase transition from VO_2(M) to VO_2(R). *J. Appl. Phys.* **110,** 113517 (2011).
22. Hellman, O., Steneteg, P., Abrikosov, I. A. & Simak, S. I. Temperature dependent effective potential method for accurate free energy calculations of solids. *Phys. Rev. B* **87,** 104111 (2013).
23. Comes, R., Felix, P., Lambert, M. & Villeneuve, G. Metal to insulator phase-transition in $V_{0.90}Nb_{0.10}O_2$ explained by local pairing of vanadium atoms. *Acta Crystallogr. A* **30,** 55–60 (1974).
24. Traylor, J. G., Smith, H. G., Nicklow, R. M. & Wilkinson, M. K. Lattice dynamics of rutile. *Phys. Rev. B* **3,** 3457–3472 (1971).
25. Haverkort, M. W. *et al.* Orbital-assisted metal-insulator transition in VO_2. *Phys. Rev. Lett.* **95,** 196404 (2005).
26. Khomskii, D. I. & Mizokawa, T. Orbitally induced Peierls state in spinels. *Phys. Rev. Lett.* **94,** 156402 (2005).
27. Tomczak, J. M., Aryasetiawan, F. & Biermann, S. Effective bandstructure in the insulating phase versus strong dynamical correlations in metallic VO_2. *Phys. Rev. B* **78,** 115103 (2008).
28. Weber, C. *et al.* Vanadium dioxide: a Peierls-Mott insulator stable against disorder. *Phys. Rev. Lett.* **108,** 256402 (2012).
29. Yuan, X., Zhang, Y. B., Abtew, T. A., Zhang, P. H. & Zhang, W. Q. VO_2: Orbital competition, magnetism, and phase stability. *Phys. Rev. B* **86,** 235103 (2012).
30. Eyert, V. VO_2: a novel view from band theory. *Phys. Rev. Lett.* **107,** 016401 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements Research by J.D.B., O.D., M.E.M., E.D.S., L.A.B. and R.J.M. was supported by the US Department of Energy (DOE), Basic Energy Sciences (BES), Materials Sciences and Engineering Division (MSED). Research by J.H. was supported by the Center for Accelerating Materials Modeling, funded by the US DOE, BES, MSED. Experimental work by C.W.L. was sponsored by the Laboratory Directed Research and Development Program of ORNL (Principal Investigator, O.D.). Research by D.L.A. at the Spallation Neutron Source and J.Z.T., A.H.S. and B.M.L. at the Advanced Photon Source (APS), Argonne National Laboratory (ANL), was supported by the US DOE, BES, Scientific User Facilities Division. We thank A. Tselev, S. Nagler, A. Banerjee, H. Krakauer and V. Cooper for interesting discussions on VO_2. Inelastic neutron scattering measurements were performed using the ARCS facility at the ORNL Spallation Neutron Source, which is sponsored by the Scientific User Facilities Division, Office of Basic Energy Sciences, US Department of Energy. We thank J. Niedziela for help with the sample environment at ARCS. IXS measurements were performed using the X-ray Operations and Research (XOR) beamline 30-ID (HERIX) at the APS. Diffuse X-ray scattering measurements were performed using the XOR beamline 33-BM-C at the APS. We thank J. Karapetrova and C. Schleputz for assistance in setting up experiments at UNICAT. Use of the APS, an Office of Science User Facility operated for the US DOE Office of Science by ANL, was supported by the US DOE under contract no. DE-AC02-06CH11357. Theoretical calculations were performed using resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the US Department of Energy under contract no. DE-AC02-05CH11231. We thank O. Hellman for providing the temperature-dependent effective potential software and assistance.

Author Contributions This project included significant contributions from many researchers and all authors participated in scientific discussions. J.D.B. (experiment) and O.D. (experiment and calculations) designed this research project. L.A.B. synthesized single-crystal samples. J.H. and O.D. performed the theoretical calculations with analysis. M.E.M., C.W.L., J.D.B., O.D. and D.L.A. performed the INS measurements and analysis. E.D.S., J.D.B., O.D., C.W.L. and J.Z.T. performed the diffuse X-ray scattering measurements and analysis. J.D.B., M.E.M., O.D., C.W.L., A.H.S., B.M.L., J.Z.T. and R.J.M. performed the IXS measurements and analysis. O.D., J.D.B., M.E.M., E.D.S. and J.H. wrote the manuscript with assistance from C.W.L.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.D.B. (experiments; budaijd@ornl.gov) or O.D. (simulations; delaireoa@ornl.gov).

LETTER

doi:10.1038/nature13883

Passive radiative cooling below ambient air temperature under direct sunlight

Aaswath P. Raman[1], Marc Abou Anoma[2], Linxiao Zhu[3], Eden Rephaeli[1] & Shanhui Fan[1]

Cooling is a significant end-use of energy globally and a major driver of peak electricity demand. Air conditioning, for example, accounts for nearly fifteen per cent of the primary energy used by buildings in the United States[1]. A passive cooling strategy that cools without any electricity input could therefore have a significant impact on global energy consumption. To achieve cooling one needs to be able to reach and maintain a temperature below that of the ambient air. At night, passive cooling below ambient air temperature has been demonstrated using a technique known as radiative cooling, in which a device exposed to the sky is used to radiate heat to outer space through a transparency window in the atmosphere between 8 and 13 micrometres[2–11]. Peak cooling demand, however, occurs during the daytime. Daytime radiative cooling to a temperature below ambient of a surface under direct sunlight has not been achieved[3,4,12,13] because sky access during the day results in heating of the radiative cooler by the Sun. Here, we experimentally demonstrate radiative cooling to nearly 5 degrees Celsius below the ambient air temperature under direct sunlight. Using a thermal photonic approach[14–25], we introduce an integrated photonic solar reflector and thermal emitter consisting of seven layers of HfO_2 and SiO_2 that reflects 97 per cent of incident sunlight while emitting strongly and selectively in the atmospheric transparency window. When exposed to direct sunlight exceeding 850 watts per square metre on a rooftop, the photonic radiative cooler cools to 4.9 degrees Celsius below ambient air temperature, and has a cooling power of 40.1 watts per square metre at ambient air temperature. These results demonstrate that a tailored, photonic approach can fundamentally enable new technological possibilities for energy efficiency. Further, the cold darkness of the Universe can be used as a renewable thermodynamic resource, even during the hottest hours of the day.

Consider a radiative cooler of area A at temperature T, whose spectral and angular emissivity is $\epsilon(\lambda,\theta)$. When the radiative cooler is exposed to a daylight sky, it is subject to both solar irradiance and atmospheric thermal radiation (corresponding to ambient air temperature T_{amb}). The net cooling power P_{cool} of such a radiative cooler is given by:

$$P_{cool}(T) = P_{rad}(T) - P_{atm}(T_{amb}) - P_{Sun} - P_{cond+conv} \quad (1)$$

In equation (1) the power radiated out by the structure is:

$$P_{rad}(T) = A\int d\Omega \cos\theta \int_0^\infty d\lambda I_{BB}(T,\lambda)\epsilon(\lambda,\theta) \quad (2)$$

Here $\int d\Omega = 2\pi\int_0^{\pi/2} d\theta \sin\theta$ is the angular integral over a hemisphere. $I_{BB}(T,\lambda) = \frac{2hc^2}{\lambda^5}\frac{1}{e^{hc/(\lambda k_B T)}-1}$ is the spectral radiance of a blackbody at temperature T, where h is Planck's constant, k_B is the Boltzmann constant, c is the speed of light and λ is the wavelength.

$$P_{atm}(T_{amb}) = A\int d\Omega \cos\theta \int_0^\infty d\lambda I_{BB}(T_{amb},\lambda)\epsilon(\lambda,\theta)\epsilon_{atm}(\lambda,\theta) \quad (3)$$

is the absorbed power due to incident atmospheric thermal radiation, and:

$$P_{Sun} = A\int_0^\infty d\lambda \epsilon(\lambda,\theta_{Sun})\, I_{AM1.5}(\lambda) \quad (4)$$

is the incident solar power absorbed by the structure. We arrive at equation (3) and equation (4) by using Kirchhoff's radiation law to replace the structure's absorptivity with its emissivity $\epsilon(\lambda,\theta)$. The angle-dependent emissivity of the atmosphere is given by[6]: $\epsilon_{atm}(\lambda,\theta) = 1 - t(\lambda)^{1/\cos\theta}$, where $t(\lambda)$ is the atmospheric transmittance in the zenith direction[26]. In equation (4), the solar illumination is represented by $I_{AM1.5}(\lambda)$, the AM1.5 spectrum. We assume the structure is facing the Sun at a fixed angle θ_{Sun}. Thus the term P_{Sun} does not have an angular integral, and the structure's emissivity is represented by its value at θ_{Sun}.

$$P_{cond+conv}(T,T_{amb}) = Ah_c(T_{amb} - T) \quad (5)$$

is the power lost due to convection and conduction. $h_c = h_{cond} + h_{conv}$ is a combined non-radiative heat coefficient that captures the collective effect of conductive and convective heating owing to the contact of the radiative cooler with external surfaces and air adjacent to the radiative cooler.

Equation (1) in general relates the cooling power $P_{cool}(T)$ of the surface, that is, the net power outflow of the surface, as a function of its temperature. Such a surface becomes a daytime cooling device if there is a net positive power outflow when $T = T_{amb}$ under direct sunlight, that is, if it radiates more heat out to space than it gains by absorbing sunlight and atmospheric thermal radiation. The power outflow $P_{cool}(T = T_{amb})$ then defines its cooling power at ambient air temperature. In the absence of net outflow, a radiative cooler's temperature should reach a steady-state temperature below ambient. The solution of equation (1) with $P_{cool}(T) = 0$ defines the steady-state temperature T_s. The goal of our experiment is to demonstrate a daytime radiative cooling device with $T_s < T_{amb}$, and to measure its cooling power as a function of T under direct sunlight, corresponding to peak daytime conditions.

To achieve daytime radiative cooling, the device must satisfy a very stringent set of constraints as dictated by the power balance equation of equation (1). First, it must reflect sunlight strongly to minimize P_{Sun}. Therefore, it must be strongly reflecting over visible and near-infrared wavelength ranges. Second, it must strongly emit thermal radiation P_{rad} while minimizing incident atmospheric thermal radiation P_{atm} by minimizing its emission at wavelengths where the atmosphere is opaque. Thus, the device must emit selectively and strongly only between 8 μm and 13 μm, where the atmosphere is transparent, and reflect at all other wavelengths. These constraints are formidable and fundamentally thermodynamic in nature. Radiative power scales as T^4, and the Sun, at 5,777 K, far outstrips the radiation of room-temperature objects on Earth, which are typically around 300 K. Even with an ideally selective emitter that emits only in the atmospheric transparency window, over 90% of incident sunlight must be reflected to remain at ambient temperature. In practice, to achieve meaningful daytime radiative cooling more than

[1]Ginzton Laboratory, Department of Electrical Engineering, Stanford University, Stanford, California 94305, USA. [2]Department of Mechanical Engineering, Stanford University, Stanford, California 94305, USA. [3]Department of Applied Physics, Stanford University, Stanford, California 94305, USA.

94% of sunlight must be reflected, especially given variation in atmospheric conditions across different geographic regions[27]. This is particularly challenging when combined with the goal of emitting strongly and selectively in the atmospheric window. Previous approaches using metallic reflectors and conventional thermal emitters with reflective cover foils have thus proved to be insufficient to achieve cooling under direct sunlight. Finally, the radiative cooler must be well sealed from its environment to minimize h_c and in turn $P_{cond+conv}$. This constraint presents an experimental design challenge during the daytime given that most surfaces that might be in contact with the radiative cooler will themselves heat up when exposed to sunlight and transfer this added heat to the cooler.

A previous paper presented a theoretical design of a photonic structure capable of satisfying the emission and reflection requirements for cooling[14]. The design there involved the use of a complex two-dimensional photonic crystal that would require photolithography. Here we introduce and numerically optimize an alternative theoretical design based on one-dimensional photonic films that is more amenable to large-scale fabrication, and experimentally realize it. Furthermore, we design and build an apparatus that minimizes heat load on the radiative cooler, allowing us to observe below-ambient cooling in the daytime for the first time.

The rooftop measurement apparatus that minimizes h_c, and is used to experimentally demonstrate radiative cooling under direct sunlight, is shown in Fig. 1a. The design of the apparatus, shown schematically in Fig. 1b and c, reduces both convection and conduction to the radiative cooler under peak solar irradiance. The radiative cooling surface, deposited on a 200-mm silicon wafer, is placed on a polystyrene pedestal which is supported by a clear acrylic box (see Methods for details). A clear 12.5-μm polyethylene film lies above the sample as an infrared-transparent wind shield. As can be seen in the two-dimensional schematic of the apparatus in Fig. 1c, the radiative cooler is thus suspended in a relatively well-sealed air pocket. Such an air pocket design represents a key innovation in the experimental demonstration of daytime radiative cooling: note that our sample is far smaller than the surrounding roof. This design aims to ensure that any surface in immediate contact with the air pocket or the sample will heat up minimally due to solar irradiance. Finally, to ensure peak sunlight irradiance of up to 890 W m^{-2} on the radiative cooler during the winter months in which testing was conducted, the entire apparatus is tilted 30° towards the south. This experimental constraint reduces sky access for the purposes of thermal radiation, so for the same setup, better cooling performance would be expected if one were to operate the cooler without the tilt.

We use a photonic approach to meet these stringent demands of selective thermal emission in the mid-infrared, and strong solar reflection. An extensive numerical optimization scheme (see Methods) is used to achieve the photonic design schematically shown in Fig. 1d with a scanning electron microscope cross-section. The photonic radiative cooler consists of seven alternating layers of hafnium dioxide (HfO_2) and silicon dioxide (SiO_2) of varying thicknesses, on top of 200 nm of silver (Ag), which are all deposited on top of a 200-mm silicon wafer. The bottom four layers of HfO_2 and SiO_2 have thicknesses that are less than 100 nm and assist in optimizing solar reflection in a manner akin to that achievable using periodic one-dimensional photonic crystals. HfO_2 serves as a high-index material that also presents low ultraviolet absorption, a useful feature when optimizing for solar reflectance, while SiO_2 is optically transparent and is the low-index layer. The use of HfO_2 is, however, not essential, and can be replaced with titanium dioxide (TiO_2), which is less expensive. The top three layers are much thicker and are primarily responsible for thermal radiation from the cooler,

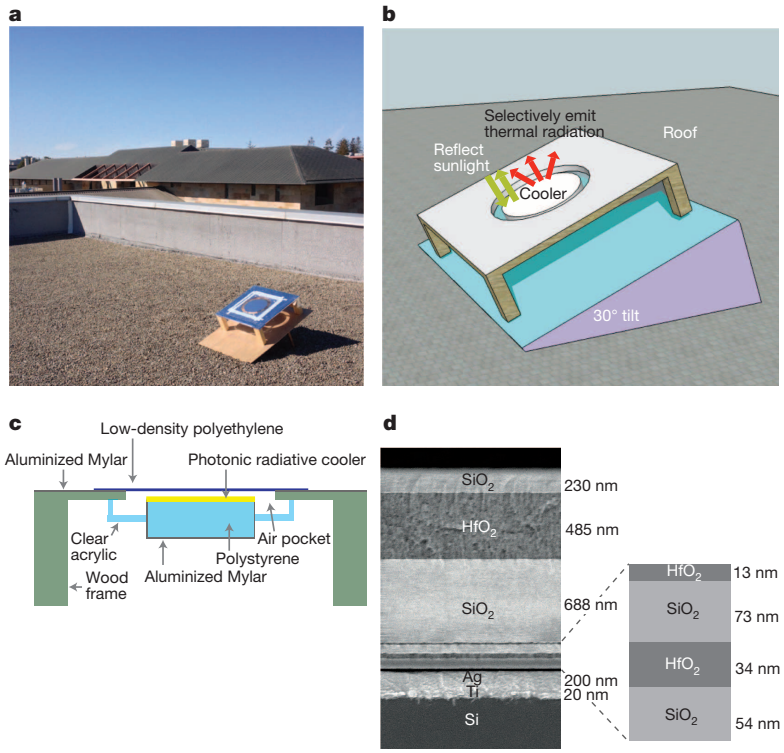


Figure 1 | Rooftop apparatus and photonic radiative cooler. **a**, Photo of the apparatus and radiative cooler on the test rooftop in Stanford, California. **b**, Three-dimensional schematic of the apparatus and radiative cooler, showing the general mode of operation of the radiative cooler. The apparatus is designed to minimize conductive and convective heat exchange to the cooler. **c**, Cut-out schematic of the apparatus through the middle, showing how an air pocket is created around the radiative cooler. Surfaces adjacent to this air pocket heat up minimally due to incident solar irradiance and therefore minimize the heat load on the air inside the pocket. Mylar is polyethylene terephthalate. **d**, Scanning electron microscope image of the photonic radiative cooler that is designed, implemented and tested in our experiments. It consists of seven layers of HfO_2 and SiO_2, whose thicknesses are defined by extensive numerical optimization (see Methods), on top of 200 nm of Ag, a 20-nm-thick Ti adhesion layer, and a 750-μm-thick, 200-mm-diameter Si wafer substrate.

through a combination of material properties and interference effects. SiO_2 has a strong peak in its absorptivity near 9 μm due to its phonon–polariton resonance. HfO_2 also presents non-zero absorption and hence emission in the 8–13 μm wavelength range[28]. The combination of all these layers results in a macroscopically planar and integrated structure that collectively achieves high solar reflectance and strong thermal emission.

The photonic radiative cooler's absorptivity/emissivity spectrum is experimentally characterized and shown in Fig. 2. The cooler shows minimal absorption when integrated from 300 nm to 4 μm, where the solar spectrum is present, in Fig. 2a, reflecting 97% of incident solar power at near-normal incidence. In Fig. 2b we observe that the cooler has strong and remarkably selective emissivity in the atmospheric window between 8 μm and 13 μm. Moreover, the photonic radiative cooler's thermal emissivity persists to large angles (see Extended Data Fig. 1), a useful feature to maximize radiated power P_{rad}, a hemispherically integrated quantity—see equation (2)—and reminiscent of the behaviour of hyperbolic metamaterials[29]. Photonic design fundamentally enables these spectral properties, which in turn are essential to achieving below-ambient radiative cooling. This spectral behaviour, and below-ambient cooling, is not achievable using these materials individually with conventional metallic reflectors.

We demonstrate the performance of the photonic radiative cooler on a clear winter day in Stanford, California, by exposing it to the sky on a building roof during daylight hours and comparing its steady-state temperature to the ambient air temperature. As shown in the temperature data of Fig. 3a, immediately after the sample is exposed to the environment (shortly before 10:00 local time in Fig. 3a), its temperature drops to approximately 4° to 5° Celsius below the ambient air temperature, even though significant solar irradiance is already incident on the sample. This is a key signature of radiative cooling, and a counterintuitive result during the day: we typically think of surfaces increasing their temperature when removed from the shade and exposed to the Sun during the day. We observe the photonic radiative cooler's temperature for over five hours under direct sunlight. Over 800 W m^{-2} of solar power is incident on the sample for three of the five hours. The cooler maintains a steady-state temperature substantially below the air temperature over the entire day, and is 4.9°C ± 0.15°C below the air temperature between 13:00 and 14:00 (local time) when the solar irradiance is in the range 800–870 W m^{-2}. To illustrate the significance of this result, we compare in Fig. 3b the photonic radiative cooler's performance against 200-mm wafers in identical apparatuses coated with conventional materials: carbon black paint and aluminium. The black paint reaches near 80° C, which is more than 60 °C above the ambient air temperature, while the aluminium reaches nearly 40° C, which is 20° C above the ambient air temperature. Typical roofing material has strong solar absorption and hence significantly heats up under direct sunlight, as emulated by the black paint result here. Also, one still sees very strong heating with an aluminium film, even though it provides relatively strong solar reflection.

We next characterize the photonic radiative cooler's cooling power. We allow its temperature to reach the previously achieved steady-state value under peak sunlight conditions of nearly 900 W m^{-2}. We then input heat to the cooler in steps over the course of one hour and observed

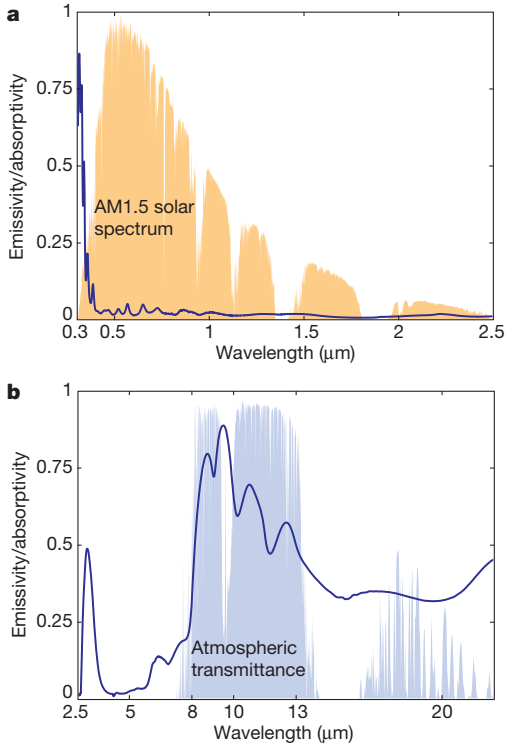


Figure 2 | Emissivity/absorptivity of the photonic radiative cooler from the ultraviolet to the mid-infrared. **a**, Measured emissivity/absorptivity at 5° angle of incidence of the photonic radiative cooler over optical and near-infrared wavelengths using an unpolarized light source, with the AM1.5 solar spectrum plotted for reference. The cooler reflects 97% of incident solar radiation. **b**, Measured emissivity/absorptivity of the cooler at 5° angle of incidence over mid-infrared wavelengths using an unpolarized light source, with a realistic atmospheric transmittance model plotted for reference[26]. The photonic cooler achieves strong selective emission within the atmospheric window.

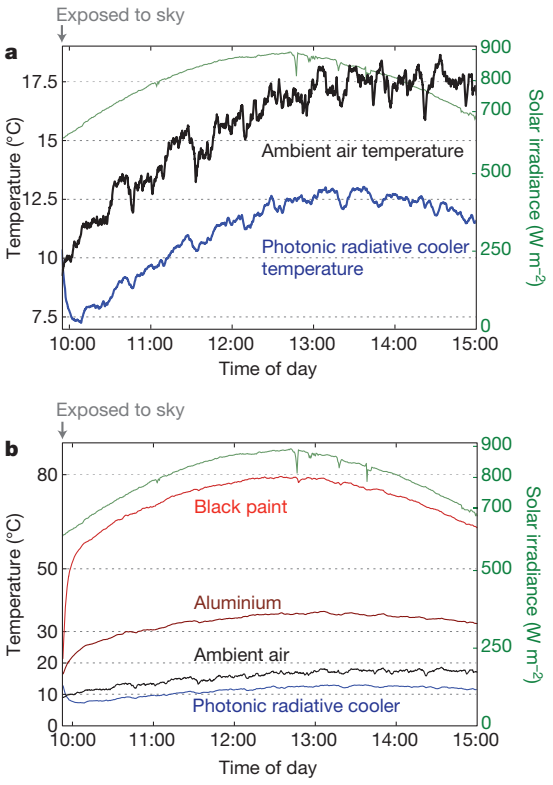


Figure 3 | Steady-state temperature of photonic radiative cooler. **a**, Rooftop measurement of the photonic radiative cooler's performance (blue) against ambient air temperature (black) on a clear winter day in Stanford, California. The photonic radiative cooler immediately drops below ambient once exposed to the sky, and achieves a steady-state temperature T_s of 4.9 °C ± 0.15 °C below ambient for over one hour where the solar irradiance incident on it (green) ranges from 800 W m^{-2} to 870 W m^{-2}. **b**, Comparing the photonic radiative cooler's performance against two reference roofing materials: black paint and aluminium. The paint reaches a temperature up to 80 °C, or 60 °C above ambient, while the aluminium reaches nearly 40 °C, or 20 °C above ambient. Only the photonic cooler stays well below ambient under direct solar irradiance.

the cooler's temperature at each step, as shown in Fig. 4a. With each increase of heat input, the temperature of the cooler rises to a new steady state. We plot the temperature of the cooler as a function of heat power in Fig. 4b. The temperature of the cooler reaches ambient temperature with an input heat power of 40.1 ± 4.1 W m^{-2}, indicating that substantial cooling power is available from this device. We next develop a theoretical model of our photonic cooler. This model is based on equation (1), where we use the spectral data of Fig. 2, as well as a model of the atmospheric transmittance[26] (see Methods), and a model for the conductive and convective losses of the apparatus, that together yield a value of $h_c = 6.9\ \mathrm{W\ m^{-2}\ K^{-1}}$ (see Methods and Extended Data Fig. 2). The theoretical model agrees excellently with the experimental data (Fig. 4b). The model can also be used to predict the steady-state temperature and power balance of the photonic radiative cooler as a function of time, and as compared against the observed performance of the sample under both daytime and night-time conditions (Extended Data Fig. 3). Moreover, the model indicates that lower steady-state temperatures can be reached by the cooler by reducing convective losses (see Methods). Under the same atmospheric and solar conditions, but with $h_c \to 0$, our device should achieve a steady-state temperature of 19.5 °C below ambient (see Extended Data Fig. 2c). Substantial gains in the photonic radiative cooler's performance are thus achievable by improved packaging alone.

The below-ambient cooling under peak daylight conditions shown here presents the possibility for purely passive, water-free approaches to cooling buildings and vehicles at all hours of the day. A preliminary analysis indicates that photonic radiative coolers could compete favourably in economic terms against other rooftop renewable energy options for cooling such as photovoltaic panels but may also work cooperatively with them (see Methods and Extended Data Fig. 4). A key engineering challenge will be to minimize parasitic heating of the radiative cooler from the surroundings while delivering the desired heat load to it efficiently. With the remarkable degree of spectral and thermal control made possible by photonic approaches, cooling power performance in favourable atmospheric conditions could be improved to more than 100 W m^{-2} (ref. 14). Improving building efficiency with a view towards reducing the need for active cooling has taken on renewed urgency on our warming planet, where the increase in carbon emissions caused by air conditioning is predicted to be faster than the decline in emissions from reduced heating[30]. In off-grid areas of the developing world, achieving radiative cooling during the daytime offers the opportunity to enable electricity-free cooling for critical needs like long-term food and medical supply storage. More broadly, our results point to the largely unexplored opportunity of using the cold darkness of the Universe as a fundamental renewable thermodynamic resource for improving energy efficiency here on Earth.

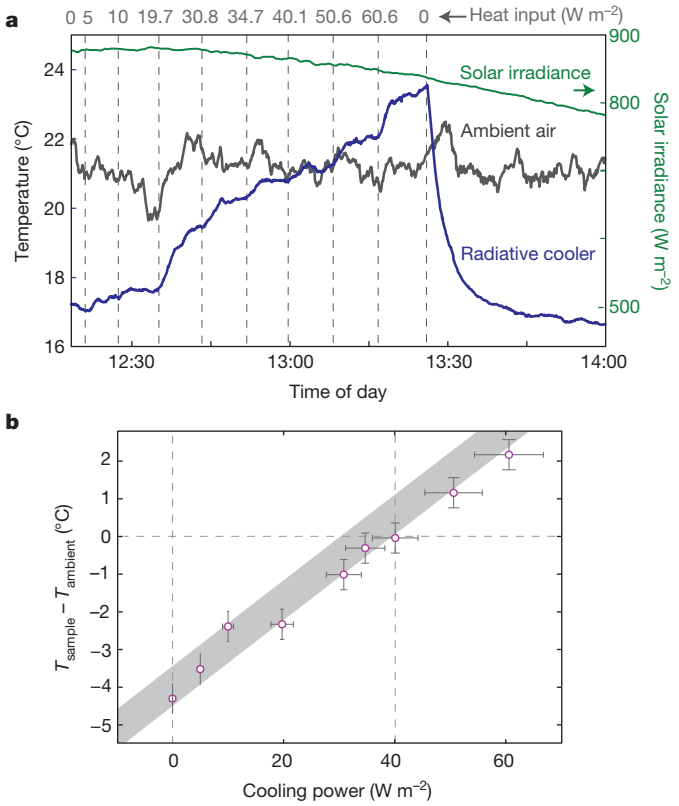


Figure 4 | Cooling power of photonic radiative cooler. **a**, Rooftop measurement of the photonic radiative cooler's temperature (blue), relative to ambient air temperature (black), in response to step-wise increasing inputs of heat (quantity shown at top at the beginning of each time period). For an input of 40.1 ± 4.1 W m^{-2} the radiative cooler reaches ambient air temperature, where solar irradiance of 860 W m^{-2} is incident on the cooler. **b**, The average temperature reached by the cooler over the last two minutes of each time period in **a** relative to ambient air temperature (T_{sample} minus $T_{ambient}$) for each associated value of input heat, with error bars defined by instrument and measurement error (see Methods). The predictions of a theoretical model for $h_c = 6.9\ \mathrm{W\ m^{-2}\ K^{-1}}$ are shown in the grey band, whose bounds are set by uncertainty associated with atmospheric transmittance.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 23 April; accepted 22 September 2014.

1. Kelso, J. K. (ed.) *2011 Buildings Energy Data Book* http://buildingsdatabook.eren.doe.gov/docs/DataBooks/2011_BEDB.pdf (US Department of Energy, Office of Energy Efficiency and Renewable Energy, 2011).
2. Trombe, F. Perspectives sur l'utilisation des rayonnements solaires et terrestres dans certaines régions du monde. *Revue Générale Thermique* **6,** 1285–1314 (1967).
3. Catalanotti, S. *et al.* The radiative cooling of selective surfaces. *Sol. Energy* **17,** 83–89 (1975).
4. Bartoli, B. *et al.* Nocturnal and diurnal performances of selective radiators. *Appl. Energy* **3,** 267–286 (1977).
5. Granqvist, C. G. & Hjortsberg, A. Surfaces for radiative cooling: silicon monoxide films on aluminum. *Appl. Phys. Lett.* **36,** 139–141 (1980).
6. Granqvist, C. G. & Hjortsberg, A. Radiative cooling to low temperatures: general considerations and application to selectively emitting SiO films. *J. Appl. Phys.* **52,** 4205–4220 (1981).
7. Berdahl, P., Martin, M. & Sakkal, F. Thermal performance of radiative cooling panels. *Int. J. Heat Mass Transf.* **26,** 871–880 (1983).
8. Berdahl, P. Radiative cooling with MgO and/or LiF layers. *Appl. Opt.* **23,** 370–372 (1984).
9. Orel, B., Gunde, M. & Krainer, A. Radiative cooling efficiency of white pigmented paints. *Sol. Energy* **50,** 477–482 (1993).
10. Gentle, A. R. & Smith, G. B. Radiative heat pumping from the earth using surface phonon resonant nanoparticles. *Nano Lett.* **10,** 373–379 (2010).
11. Gentle, A., Aguilar, J. & Smith, G. Optimized cool roofs: integrating albedo and thermal emittance with *R*-value. *Sol. Energy Mater. Sol. Cells* **95,** 3207–3215 (2011).
12. Nilsson, T. M. & Niklasson, G. A. Radiative cooling during the day: simulations and experiments on pigmented polyethylene cover foils. *Sol. Energy Mater. Sol. Cells* **37,** 93–118 (1995).
13. Nilsson, T. M., Niklasson, G. A. & Granqvist, C. G. A solar reflecting material for radiative cooling applications: ZnS pigmented polyethylene. *Sol. Energy Mater. Sol. Cells* **28,** 175–193 (1992).
14. Rephaeli, E., Raman, A. & Fan, S. Ultrabroadband photonic structures to achieve high-performance daytime radiative cooling. *Nano Lett.* **13,** 1457–1461 (2013).
15. Lin, S.-Y. *et al.* Enhancement and suppression of thermal emission by a three-dimensional photonic crystal. *Phys. Rev. B* **62,** R2243–R2246 (2000).
16. Greffet, J.-J. *et al.* Coherent emission of light by thermal sources. *Nature* **416,** 61–64 (2002).
17. Narayanaswamy, A. & Chen, G. Thermal emission control with one-dimensional metallodielectric photonic crystals. *Phys. Rev. B* **70,** 125101 (2004).
18. Luo, C., Narayanaswamy, A., Chen, G. & Joannopoulos, J. D. Thermal radiation from photonic crystals: a direct calculation. *Phys. Rev. Lett.* **93,** 213905 (2004).
19. Lee, B. J., Fu, C. J. & Zhang, Z. M. Coherent thermal emission from one-dimensional photonic crystals. *Appl. Phys. Lett.* **87,** 071904 (2005).
20. Drevillon, J. & Ben-Abdallah, P. Ab initio design of coherent thermal sources. *J. Appl. Phys.* **102,** 114305 (2007).
21. Schuller, J., Taubner, T. & Brongersma, M. Optical antenna thermal emitters. *Nature Photon.* **3,** 658–661 (2009).
22. Rephaeli, E. & Fan, S. Absorber and emitter for solar thermo-photovoltaic systems to achieve efficiency exceeding the Shockley-Queisser limit. *Opt. Express* **17,** 15145–15159 (2009).

23. Wu, C. *et al.* Metamaterial-based integrated plasmonic absorber/emitter for solar thermo-photovoltaic systems. *J. Opt.* **14,** 024005 (2012).
24. De Zoysa, M. *et al.* Conversion of broadband to narrowband thermal emission through energy recycling. *Nature Photon.* **6,** 535–539 (2012).
25. Lenert, A. *et al.* A nanophotonic solar thermophotovoltaic device. *Nature Nanotechnol.* **9,** 126–130 (2014).
26. Berk, A. *et al.* Modtran5: 2006 update. *Proc. SPIE* **6233,** 62331F (2006).
27. Martin, M. & Berdahl, P. Characteristics of infrared sky radiation in the United States. *Sol. Energy* **33,** 321–336 (1984).
28. Bright, T., Watjen, J., Zhang, Z., Muratore, C. & Voevodin, A. Optical properties of HfO_2 thin films deposited by magnetron sputtering: from the visible to the far-infrared. *Thin Solid Films* **520,** 6793–6802 (2012).
29. Jacob, Z. *et al.* Engineering photonic density of states using metamaterials. *Appl. Phys. B* **100,** 215–218 (2010).
30. Isaac, M. & van Vuuren, D. P. Modeling global residential sector energy demand for heating and air conditioning in the context of climate change. *Energy Policy* **37,** 507–521 (2009).

Acknowledgements This work is supported by the Advanced Research Projects Agency-Energy (ARPA-E), Department of Energy (contract number DE-AR0000316). We acknowledge discussions with J. Eaton and K. Goodson. Part of this work was performed at the Stanford Nanofabrication Facility, which is supported by the National Science Foundation through the NNIN under grant number ECS-9731293, and the Stanford Nano Center (SNC)/Stanford Nanocharacterization Laboratory (SNL), part of the Stanford Nano Shared Facilities.

Author Contributions A.P.R. and S.F. envisioned and implemented the experimental studies, and wrote the manuscript. A.P.R. and M.A.A. built and executed the rooftop experiments. A.P.R. designed and characterized the radiative cooler. L.Z. and E.R. provided technical support and conceptual advice.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.P.R. (aaswath@stanford.edu) or S.F. (shanhui@stanford.edu).

LETTER

doi:10.1038/nature13951

Design and fabrication of memory devices based on nanoscale polyoxometalate clusters

Christoph Busche[1]*, Laia Vilà-Nadal[1]*, Jun Yan[1], Haralampos N. Miras[1], De-Liang Long[1], Vihar P. Georgiev[2], Asen Asenov[2], Rasmus H. Pedersen[2], Nikolaj Gadegaard[2], Muhammad M. Mirza[2], Douglas J. Paul[2], Josep M. Poblet[3] & Leroy Cronin[1]

Flash memory devices—that is, non-volatile computer storage media that can be electrically erased and reprogrammed—are vital for portable electronics, but the scaling down of metal–oxide–semiconductor (MOS) flash memory to sizes of below ten nanometres per data cell presents challenges. Molecules have been proposed to replace MOS flash memory[1], but they suffer from low electrical conductivity, high resistance, low device yield, and finite thermal stability, limiting their integration into current MOS technologies. Although great advances have been made in the pursuit of molecule-based flash memory[2], there are a number of significant barriers to the realization of devices using conventional MOS technologies[3–7]. Here we show that core–shell polyoxometalate (POM) molecules[8] can act as candidate storage nodes for MOS flash memory. Realistic, industry-standard device simulations validate our approach at the nanometre scale, where the device performance is determined mainly by the number of molecules in the storage media and not by their position. To exploit the nature of the core–shell POM clusters, we show, at both the molecular and device level, that embedding $[(Se^{IV}O_3)_2]^{4-}$ as an oxidizable dopant in the cluster core allows the oxidation of the molecule to a $[Se^{V}_2O_6]^{2-}$ moiety containing a {Se(v)–Se(v)} bond (where curly brackets indicate a moiety, not a molecule) and reveals a new 5+ oxidation state for selenium. This new oxidation state can be observed at the device level, resulting in a new type of memory, which we call 'write-once-erase'. Taken together, these results show that POMs have the potential to be used as a realistic nanoscale flash memory. Also, the configuration of the doped POM core may lead to new types of electrical behaviour[9–11]. This work suggests a route to the practical integration of configurable molecules in MOS technologies as the lithographic scales approach the molecular limit[12].

To engineer the flash memory devices, we selected a core–shell POM 'Dawson-like' archetype as the functional part of the switching node with the general formula $[M_{18}O_{54}(XO_n)_2]^{m-}$ (where M = Mo or W, X = P, S or Se, $n = 3$ or 4 and $m = 2 \rightarrow 8$). This is because the cluster has a nanoscale size (about $1.2 \times 1 \times 1$ nm), a wide accessible charge range to act as ideal trapped charges for flash memory, a configurable cluster core (that is, P, S, Se as possible dopants) and a high thermal stability (about 600 °C) to cope with the high temperatures associated with flash memory post-processing. With the thermal stability in mind we therefore used a tungsten-based rather than molybdenum-based core–shell cluster. Furthermore, the effect of various dopants was investigated by density functional theory (DFT) calculations, which helped us to understand the likely reactivity and electronic structure of the clusters[13], and hence to choose an appropriate heteroatom. After careful consideration of the synthetically realistic options we hypothesized that $\{Se^{IV}O_3\}$ would provide the right balance of structural stability and electronic activity, leading us to conclude that the cluster anion $[W_{18}O_{54}(SeO_3)_2]^{4-}$ would be an ideal candidate to investigate the development of practical flash memory devices.

To explore this idea we set about synthesizing the core–shell cluster anion $[W_{18}O_{54}(SeO_3)_2]^{4-}$ (**1a**) via a dehydration reaction of the selenite containing cluster $[W_{18}O_{56}(SeO_3)_2(H_2O)_2]^{8-}$ (**precursor**) (see Supplementary Information). Notably, this cluster demonstrates exceptionally rich redox behaviour associated not only with the reduction of the metal oxide cluster $\{W_{18}O_{54}\}$ cage, but also with the oxidation of selenite templates at the cluster core; see Fig. 1.

To provide a first demonstration of a flash memory cell using POMs, a lateral geometry was used with a ~4-nm Si nanowire channel covered with a 4-nm SiO_2 insulator. This design, rather than a vertical flash memory cell, was chosen to provide easy access for the exploration of the intrinsic ability of the POM to form the switching component of a flash memory device (see Fig. 2). Nominally identical flash memory characteristics were demonstrated in an array of nine independent devices. We tested devices with different geometries and only those with a distance between the control gate and the nanowire channel of below 60 nm demonstrated reproducible flash memory behaviour. The fabrication process is described in full in the Supplementary Information.

Figure 2b demonstrates a shift in the threshold voltage (ΔV_T) of 1.1 ± 0.1 V from the bare nanowire to the same nanowire coated in POMs. A large negative voltage of –20 V was then used to inject charge into the POMs before the control gate was again swept to demonstrate drain current characteristics with a ΔV_T shift of 1.2 ± 0.1 V at low voltages. After a further pulse of +20 V, the drain current characteristics returned to the characteristics close to the original uncharged state. The present device geometry is not optimized, accounting for the high voltages

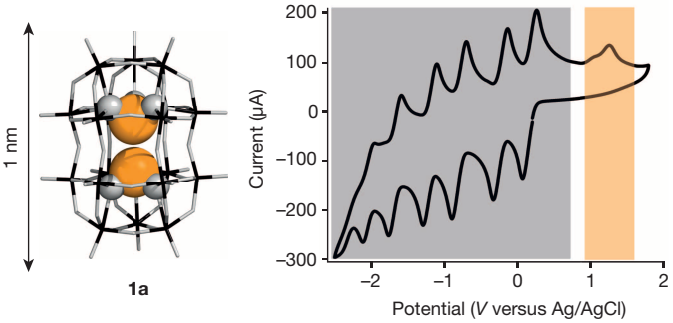


Figure 1 | Structure and electrochemical properties of compound 1a. On the left, the crystal structure of the core–shell cluster $[W_{18}O_{54}(SeO_3)_2]^{4-}$ (**1a**) is shown, with the $\{W_{18}O_{54}\}$ cage shown as black and grey lines. The two Se core dopants are shown as orange spheres. The cluster cage can be reduced multiple times (grey area) and the two Se dopants at the POM cluster core can be oxidized (orange area). On the right, the cyclic voltammetry is obtained from microcrystals of **1a** adhered to a glassy carbon electrode (diameter 1.5 mm) in 0.1 M tetrabutylammonium PF_6 acetonitrile solution at a scan rate of 200 mV s^{-1} and a scanning range V of −2.5 V to 1.8 V against a Ag/AgCl reference.

[1]WestCHEM, School of Chemistry, The University of Glasgow, Glasgow G12 8QQ, UK. [2]School of Engineering, The University of Glasgow, Glasgow G12 8LT, UK. [3]Departament de Química Física i Inorgànica, Universitat Rovira i Virgili, Marcel·lí Domingo street, 43007 Tarragona, Spain.
*These authors contributed equally to this work.

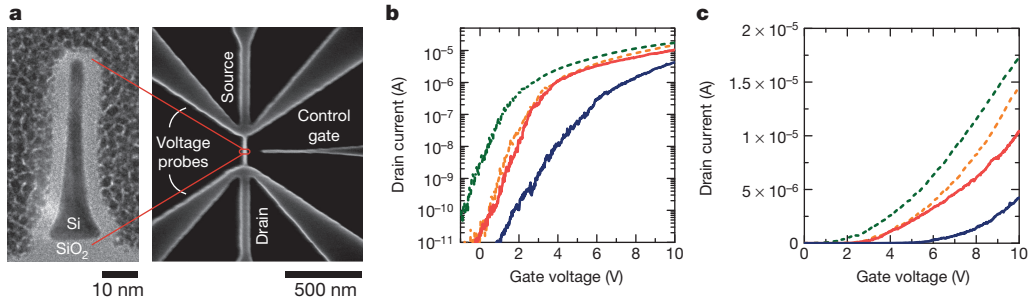


Figure 2 | Image of the flash memory device and the drain current behaviour with an applied voltage to the control gate. **a**, A cross-sectional transmission electron microscope (TEM) image (left) of the memory device with an SEM image (right) of the ~5-nm Si nanowire channel with side control gate. **b**, **c**, Measurements of the logarithmic (**b**) and linear (**c**) drain current versus gate voltage at 0.5 V source–drain voltage: before deposition of the POMs (green dashes), after the deposition of the POMs (orange dashes), after a −20 V pulse (blue line) and a +20 V pulse (red line). Panel **b** demonstrates depletion of the charge in the nanowire, resulting in a shift of the threshold voltage required to switch on the electrical conduction of the nanowire (orange dashed line) after **1a** was deposited around the nanowire. Therefore the control gate voltage required to produce the same drain current in the nanowire has moved to a higher voltage owing to the charge on the molecules of **1a**. The control gate was then used to charge and discharge the deposited molecules. −20 V was applied to the control gate, which charged the molecules around the nanowire, further increasing the control gate voltage required to produce the same drain current in the nanowire (solid blue lines in **b** and **c**). This effect could be reversed by applying a +20 V pulse to the control gate, which discharged the molecules and returned the control gate voltage for a fixed drain current to the original value with the uncharged **1a** molecules around the nanowire (red lines in **b** and **c**). The effect is repeatable, demonstrating a clear shift in the threshold voltage of the device when charged. The programming window (the threshold voltage change between the charged and uncharged **1a** molecules) is >1.2 V at low gate voltages.

required to write/erase but a plot of ΔV_T versus logarithmic time (see Supplementary Information) demonstrates that the present limit of the programme/erase times are 0.1 s and read times are 100 μs. The charge/discharge could be repeated many times and the retention time of the flash memory is at least 336 h, with the ultimate limit of the retention time expected to be significantly longer, given that no decay in the stored charge has yet been measured over the 336-h period. The read time is presently limited only by the *RC* time constant (the product of resistance

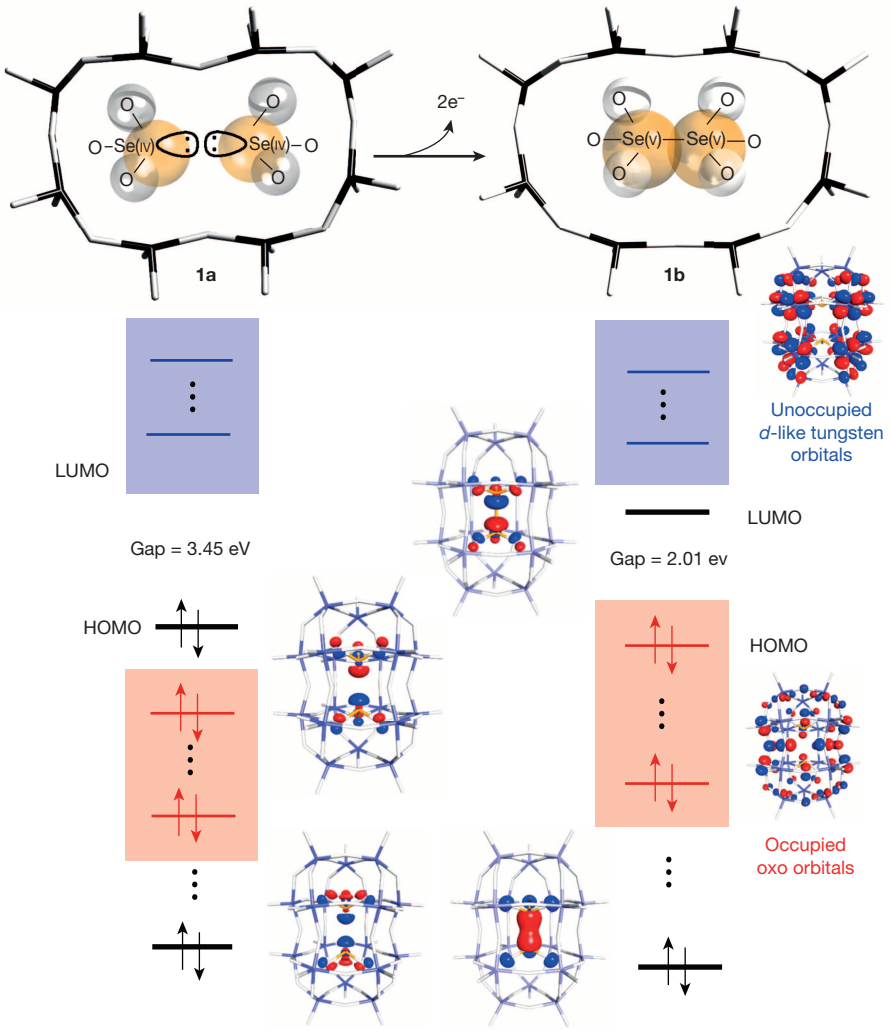
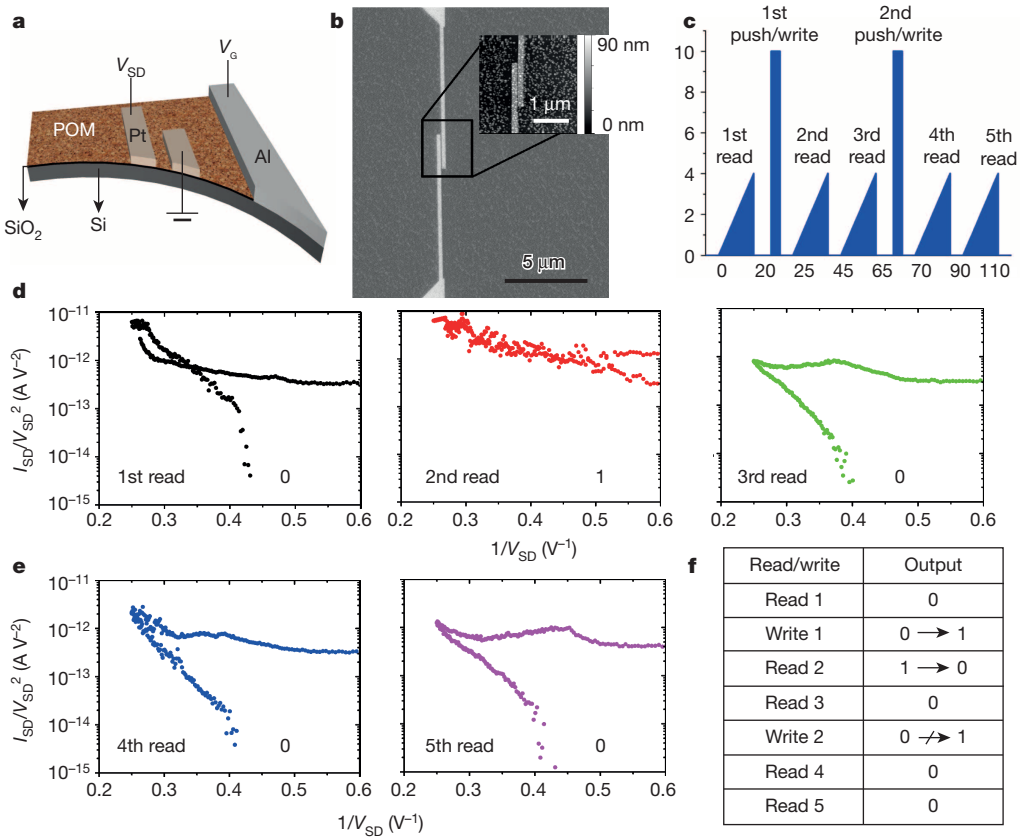


Figure 3 | Scheme depicting the formation of the Se(v)–Se(v) bond within the cluster cage. At the top, a schematic diagram shows the formation of the Se(v)–Se(v) bond in the transformation of **1a** to **1b**. At the bottom are the results from the DFT analysis, demonstrating the frontier orbitals and the formation of the Se(v)–Se(v) bond. Relevant orbitals delocalized over the Se moieties are highlighted in bold. The HOMO–LUMO gap is the energy gap between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO). Although the orbital energies of POM clusters are separated by discrete energies, they can also be viewed as having a pseudoband-like orbital structure, and in this sense the blue box depicts the set of unoccupied tungsten *d*-like orbitals and the red box the set of occupied oxygen *p*-like orbitals.

and capacitance: 22.3 pF × 250 kΩ = 56 μs) of the nanowire devices and especially by the large pad capacitance. A radio-frequency design of the device and optimization of the capacitance and resistance should reduce this to subnanosecond read times. The write/erase time is limited by the large density of POM molecules (2×10^{15} cm^{-2}) and the current compliance of the characterization set-up. A device with a shorter distance from control gate to nanowire channel and significantly fewer POMs should reduce both the write/erase voltage and the time. Preliminary calculations suggest that 100 POMs would have a subpicosecond write time, subject to the device and characterization limits, but we expect the fundamental charging mechanisms of the POM to dominate at such device dimensions. The above analysis clearly demonstrates that the ultimate performance of the POM-based flash memory has not been reached and further work is required to determine the fundamental limits of the proposed technology. The sub-threshold slope for the −20 V pulsed measurements in Fig. 2b indicates additional charging mechanisms in the device in addition to the POM flash floating-gate mechanism. Because the POMs have been distributed over the entire device with high density, there are many potential charging mechanisms that could provide this type of non-optimal behaviour. The drain current characteristics after the +20 V pulse also indicate that the return to the original state of the POMs is not complete (Fig. 2b and c), suggesting that optimization of the device geometry and POM positioning is required to improve the performance. Nevertheless, these measurements demonstrate that it is possible to produce functional flash devices using POMs owing to their intrinsic *n*-type like properties simply by drop-casting a solution of the POM directly onto the gate architecture in a one-step process.

In addition to the exploitation of the shell of the POM clusters to trap charges for functional flash memory, we also investigated the role of the two inner 'core' moieties in the POM cluster archetype $[W_{18}O_{54}(XO_n)_2]^{m-}$ using DFT (where X is P, S or Se) to see whether it is possible to use these heteroatom dopants to change the electronic structure of the cluster using a redox process. This is because, in this cluster type, the heteroatoms are perfectly positioned next to each other to interact via their lone pairs of electrons. This is particularly true for compound **1a** because the two inner {Se(IV)O_3} moieties within the outer cluster shell have significant intramolecular non-bonded interactions, with a Se(IV)···Se(IV) distance of 3.1 Å; see Fig. 3.

This was confirmed by a DFT study, which showed that the ejection of two electrons from the cluster core should lead to an oxidation state change (Se(IV) → Se(V)) in **1a** commensurate with the formation of a Se–Se bond within the cluster. Not only is this redox process revealed by the electrochemical data (see Fig. 1 and Supplementary Information), the two-electron oxidation of **1a** to **1b** was confirmed directly by coulometry (see Supplementary Information), and the formation of a Se–Se bond is consistent with studies showing that the two-electron oxidized species is diamagnetic, as confirmed by electron paramagnetic resonance spectroscopy. In addition, the theoretical estimation of the reduction potentials are in good agreement with the experimental values, as are the theoretical values for the two-electron oxidation process leading to the formation of **1b** (see Supplementary Information).



Read/write	Output
Read 1	0
Write 1	0 → 1
Read 2	1 → 0
Read 3	0
Write 2	0 ↛ 1
Read 4	0
Read 5	0

Figure 4 | The write-once-erase device. Conceptual sketch (**a**) and SEM/AFM images (**b**) of the fabricated nano-gap electrodes coated with **1a**. **c**, The measurement procedure in terms of the applied voltage. The sample was subjected to a high-voltage push pulse applied between the source and drain electrodes on the surface of the sample, and then measured at a lower voltage level. The data are obtained by sweeping the source–drain voltage (V_{SD}) between the surface electrodes from 0 V to 4 V and back to 0 V with the substrate gate voltage maintained constant at 3 V. I_{SD} is the source–drain current and V_G is the gate voltage. **d**, **e**, Fowler–Nordheim[16,17] plots of the current–voltage data of the POM-covered nano-gap electrodes, which demonstrate whether trap states which can hold charge for a memory device are present between the two source and drain electrodes (when there is hysteresis between the forward and reverse voltage sweeps, trapped charge is present). For the '0' memory state, the hysteresis in the Fowler–Nordheim plots indicates trapped charge inside the gap between the electrodes. We observe that subjecting the system to excitation with source–drain voltage at 9–10 V changes the nature of the electron transport between the source and drain upon subsequent inspection, removing the hysteresis as shown for the memory state '1'. In this measurement, the effect is transient; disappearing after the first post-excitation probe, as shown conceptually in the table in panel **f**.

To explore whether we could use the oxidative behaviour expected for the Se-embedded cluster, we fabricated a nano-electronic device incorporating **1a** to test the predictions made by the theoretical analysis. An array of parallel Pt electrodes with a gap of approximately 50 nm was produced on a highly doped silicon substrate with a thermally grown 30-nm-thick barrier oxide. Sixty-four individual electrode pairs were fabricated on each sample. A contact was opened to the silicon substrate, allowing it to function as a gate electrode. Figure 4 demonstrates scanning electron microscopy (SEM) and atomic force microscopy (AFM) images of a fabricated electrode pair, after deposition of the POM material. Through high-volume control measurements (a total of more than 250 measurements were performed on electrode pairs with and without POMs) we verified that this system is capable of probing the electrical characteristics of cluster **1b**. To do this, control experiments were undertaken with no POM material present and samples using cluster **2**, $[W_{18}O_{56}(WO_6)]^{10-}$, which has a $\{W_{18}O_{54}\}$ shell identical to that of cluster **1**, but this time contains an oxidatively inactive $\{WO_6\}$ 'core' (see Supplementary Information). These studies demonstrated that subjecting **1a** to an excitation at high source–drain bias enabled us to influence the transport characteristics upon subsequent probing at lower voltages. Figure 4 shows measurements performed with a source–drain bias of up to 4 V both before and after subjecting the system to a source–drain bias of 9–10 V. Two measurements were performed after excitation and the procedure was repeated twice, with the gate bias kept at +3 V throughout the experiment. The measurements were deliberately carried out using a slow process ensuring maximum resolution of the analyser equipment, with a full probe measurement taking approximately 20 min. Retention and dissipation of the 'write' procedure are also on this timescale and the intrinsic rate limits are expected to be similar to those of the flash memory.

Significant hysteresis is observed between the upwards and downwards voltage sweeps, with a gap of approximately 0.2 V. This is also evident in the control measurements of cluster **2** and this device thus also proves that compounds **1** and **2** are perfect examples of trapped charges giving flash-memory-like behaviour consistent with our previous observations. For the first initial post-excitation measurement however, hysteresis is not observed for compound **1a**. Absence of the hysteresis indicates a modification of the transport across and between molecules to allow easier electron flow. We therefore consider the phenomenon to be a direct representation of the electrochemically observed oxidation process Se(IV) → Se(V). This effect was observed only after the initial excitation and cannot be recreated with consecutive pushing pulses. This was typical behaviour across several devices and implies that compound **1a** can be used as a 'write-once-erase' memory. As such the positive voltage driving force for this behaviour can be directly linked to the non-reversible oxidation process of the two Se guest atoms within the POM, demonstrating how, at the device level, the molecular configuration and formation of the Se(V) dopant underpins this unprecedented behaviour.

To evaluate the possibility of incorporation of $[W_{18}O_{54}(SeO_3)_2]^{4-}$ (**1a**) and $[W_{18}O_{56}(WO_6)]^{10-}$ (**2**) molecules to achieve realization of a floating gate (see Fig. 5) in a non-volatile molecular memory, we developed a multi-scaled, multi-level computational framework designed to perform realistic flash memory cell modelling to substantially extend our previous concept[14]. Here we simulate a flash cell design with shallow trench insulation[15], which is based on a transistor with a gate length of 18 nm and gives much more accurate results, allowing us to evaluate our devices for practical implementation. Our simulation workflow links the DFT results, which are presented above, to mesoscopic (continuous) transistor simulations with the commercial three-dimensional numerical device simulator GARAND (Gold Standard Simulations Limited). The motivation for using this hierarchy of modelling approaches is the complexity of the problem. Accurate description of the POM clusters requires first-principles calculations on the atomic level, involving around 100 atoms, while the descriptions of the current flow through the flash memory cell demands continuous modelling, applied to a system of millions of atoms (see Supplementary Information).

The first step in our computational approach is to replace the poly-Si floating gate with a layer of POM molecules (Fig. 5a). More specifically, we incorporated spatial charge distributions of either a **1a** or **2** molecule for different redox states (obtained from DFT calculations) into the flash cell device structure. The POMs are negatively charged and in the native state, and those charges are counterbalanced by positively charged cations. Similar to the atomic charges, the presence of the cations

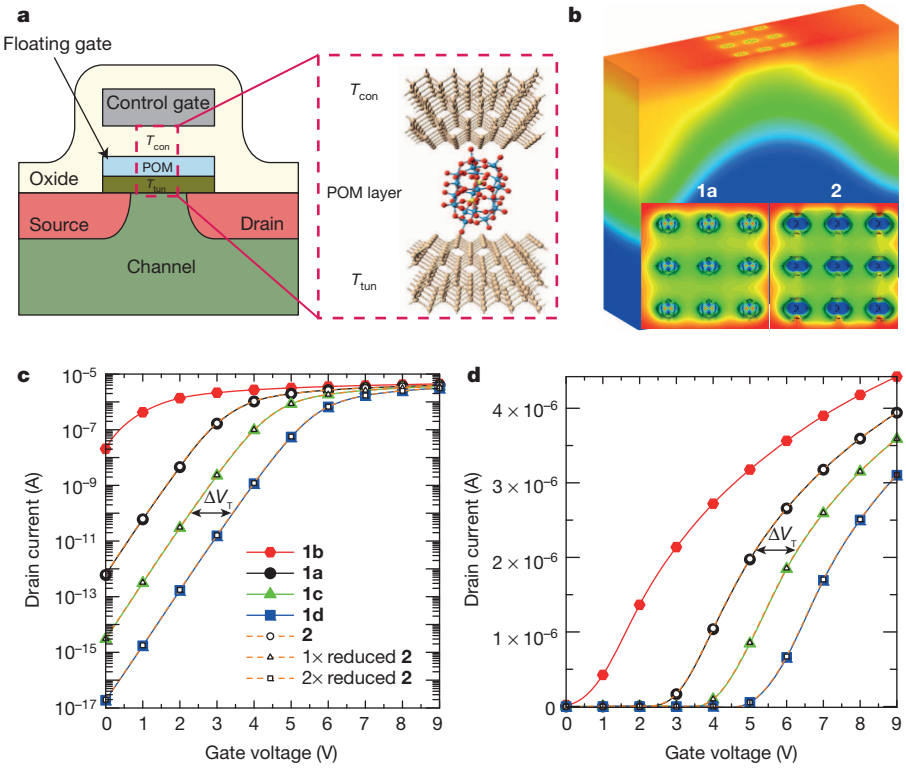


Figure 5 | Device modelling simulations of compounds 1a and 2. a, Schematic diagram representation of a single-transistor non-volatile memory cell, indicating the aimed substitution of the poly-Si floating gate with an array of POM clusters. T_{con} is the thickness of the control oxide and T_{tun} is the thickness of the tunnelling oxide. **b**, The three-dimensional electrostatic potential in the lower part of the oxide and the substrate, and two-dimensional map of the potential across the plane through the centre of the POMs, arranged in a 3 × 3 regular grid 4.5 nm from the Si–SiO_2 interface for the compounds **1a** and **2**, as schematically illustrated. **c**, **d**, Drain current versus gate voltage with a drain bias of 50 mV in logarithmic (**c**) scale and linear (**d**) scale for a bulk molecular flash cell: 2× oxidized $[W_{18}O_{54}(SeO_3)_2]^{2-}$ (**1b**), $[W_{18}O_{54}(SeO_3)_2]^{4-}$ (**1a**), 1× reduced $[W_{18}O_{54}(SeO_3)_2]^{5-}$ (**1c**) and 2× reduced $[W_{18}O_{54}(SeO_3)_2]^{6-}$ (**1d**), in comparison with $[W_{18}O_{56}(WO_6)]^{10-}$ (**2**), 1× reduced $[W_{18}O_{56}(WO_6)]^{11-}$ and 2× reduced $[W_{18}O_{56}(WO_6)]^{12-}$. V_T is the threshold voltage.

in the POM layer is modelled as a set of fractional point charges distributed around each POM. The total positive charge balances out the negative charge of the parent POMs, so that any reduction/oxidation of the POM would lead to the presence of extra electron charges in the gate stack. This provides localized balancing of each POM, essential for modelling a flash cell with broad dispersion of the position and number of POMs in the gate dielectric (see Supplementary Information).

Assuming that the POM layer consists of nine **1a** molecules arranged in a three-by-three array, we were able to evaluate the non-volatile molecular memory performance with the help of the GARAND simulator. Figure 5b shows the three-dimensional electrostatic potential in the lower part of the oxide and the substrate. In addition, the two-dimensional map of the potential across the plane through the centre of the POM layer (as arranged in a 3×3 regular grid) is presented in the same figure. From our calculations, we were able to obtain not only qualitative but also quantitative information on the impact of the oxidation/reduction of the **1a**/**2** molecular layer on the flash cell characteristics. In the process of evaluating the performance of each flash cell, the drain current I_D versus gate voltage V_G characteristic presented in Fig. 5c has an important role. Figure 5c, d shows the impact of the oxidation/reduction of the **1a**/**2** floating-gate layer on the drain current and threshold voltage of the flash memory cell. Clearly, adding electrons to the POMs leads to reduction of the OFF-current (the current value at $V_G = 0.0$ V). This is based on the fact that introducing more negative charge in the floating gate repels the electrons from the channel of the transistor. As a result, the OFF-current is reduced because it is directly influenced by the electron density distribution in the channel of the transistor (fewer electrons in the channel means less current). More importantly, the two types of POM give exactly identical I_D–V_G characteristics. The reason for this is that even though **1a** and **2** have different local charge distributions (clearly visible in the two-dimensional electrostatic potential plot in Fig. 5b), the size of each POM is very small in comparison to the channel area. Therefore, the source-to-drain current is almost unaffected by variation of the local charge distribution in the **1a** and **2** molecules. This effect is expected to increase with scaling down the channel area.

We can conclude from the results based on our multi-level computational framework that POM molecules can serve as a floating gate, with the potential for significant applications in molecular-based flash memory cells. The results demonstrate a significant programming window between each bit with a high signal-to-noise ratio. Another important characteristic for each device is the ratio of the OFF-current to the ON-current (the current at $V_G = 0.9$ V). This current ratio increases with increasing oxidation state of the POM molecule (Fig. 5c). For the clusters $[W_{18}O_{54}(SeO_3)_2]^{2-}$ (**1b**)/$[W_{18}O_{56}(WO_6)]^{10-}$ (**2**), we calculate an increase in the ratio of around two to six orders of magnitude, whereas for $[W_{18}O_{54}(SeO_3)_2]^{6-}$ (**1d**)/$[W_{18}O_{56}(WO_6)]^{12-}$, we observe a difference spanning eleven orders of magnitude.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 26 June; accepted 8 October 2014.
Published online 19 November 2014.

1. Joachim, C., Gimzewski, J. K. & Aviram, A. Electronics using hybrid-molecular and mono-molecular devices. *Nature* **408,** 541–548 (2000).
2. Shaw, J. T. *et al.* Integration of self-assembled redox molecules in flash memories. *IEEE Trans. Electron. Dev.* **58,** 826–834 (2011).
3. Zhu, H. *et al.* Non-volatile memory with self-assembled ferrocene charge trapping layer. *Appl. Phys. Lett.* **103,** 53102–53104 (2013).
4. Chen, P.-C., Shen, G. & Zhou, C. Chemical sensors and electronic noses based on 1-D metal oxide nanostructures. *IEEE Trans. Nanotechnol.* **7,** 668–682 (2008).
5. Shaw, J., Xu, Q., Rajwade, S., Hou, T.-H. & Kan, E. C. Redox molecules for a resonant tunneling barrier in nonvolatile memory. *IEEE Trans. Electron. Dev.* **59,** 1189–1198 (2012).
6. Seol, M.-L., Choi, S.-J., Kim, C.-H., Moon, D.-I. & Choi, Y.-K. Porphyrin-silicon hybrid field-effect transistor with individually addressable top-gate structure. *ACS Nano* **6,** 183–189 (2012).
7. Tans, S. J., Verschueren, A. R. M. & Dekker, C. Room-temperature transistor based on a single carbon nanotube. *Nature* **393,** 49–52 (1998).
8. Long, D. L. & Cronin, L. Towards polyoxometalate integrated nanosystems. *Chem. Eur. J.* **12,** 3698–3706 (2006).
9. Lehmann, J., Gaita-Arino, A., Coronado, E. & Loss, D. Spin qubits with electrically gated polyoxometalate molecules. *Nature Nanotechnol.* **2,** 312–317 (2007).
10. Li, H. *et al.* Layer-by-layer assembly and UV photoreduction of graphene–polyoxometalate composite films for electronics. *J. Am. Chem. Soc.* **133,** 9423–9429 (2011).
11. Fleming, L. *et al.* Surface-mediated reversible electron transfer reactions within a molecular metal oxide nano-cage. *Nature Nanotechnol.* **3,** 229–233 (2008).
12. Bonfiglio, V. & Iannaccone, G. Sensitivity-based investigation of threshold voltage variability in 32-nm flash memory cells and MOSFETs. *Solid-State Electron.* **84,** 127–131 (2013).
13. Vilà-Nadal, L. *et al.* Polyoxometalate $\{W_{18}O_{56}XO_6\}$ clusters with embedded redox-active main-group templates as localized inner-cluster radicals. *Angew. Chem. Int. Ed.* **52,** 9695–9699 (2013).
14. Vilà-Nadal, L. *et al.* Towards polyoxometalate-cluster-based nano-electronics. *Chem. Eur. J.* **19,** 16502–16511 (2013).
15. Gallon, C. *et al.* Electrical analysis of mechanical stress induced by STI in short MOSFETs using externally applied stress. *IEEE Trans. Electron. Dev.* **51,** 1254–1261 (2004).
16. Simmons, J. G. Generalized formula for the electric tunnel effect between similar electrodes separated by a thin insulating film. *J. Appl. Phys.* **34,** 1793–1803 (1963).
17. Fowler, R. H. & Nordheim, L. Electron emission in intense electric fields. *Proc. R. Soc. Lond. A* **119,** 173–181 (1928).

Supplementary Information is available in the online version of the paper.

Acknowledgements We gratefully acknowledge financial support from the EPSRC for funding (grants EP/H024107/1, EP/I033459/1 and EP/J015156/1), the COST Action CM1203 (PoCheMoN), the Royal Society Wolfson Foundation for a Merit Award, and the University of Glasgow. V.P.G. and A.A. thank S. Markov and S. M. Amoroso for discussions.

Author Contributions L.C. conceived the idea, designed the project and coordinated the efforts of the research team. J.Y. synthesised the clusters and conducted the first electrochemistry experiments and structural characterization with D.-L.L. H.N.M., C.B., L.V.-N., and L.C. helped to characterize the physical properties of the clusters. C.B. did the electron paramagnetic resonance, electrochemistry and spectroscopic measurements. L.V.-N., L.C., V.P.G. and A.A. designed the theory-to-modelling strategy. L.V.-N., with J.M.P., did the DFT calculations. V.P.G. and A.A. did the device simulation. R.H.P. and N.G. fabricated and characterized the electrode arrays, produced the devices, made the measurements and characterized the data. M.M.M. and D.J.P. designed the nanowire arrays and M.M.M. fabricated the electrodes and optimized the data with D.J.P., who helped analyse the results. C.B., L.V.-N. and L.C. co-wrote the paper with input from all the authors.

Author Information Atomic coordinates for the reported crystal structures have been deposited with the Cambridge Structural Database under the accession codes 997534 (compound **precursor**), 997535 (compound **1a**), 997536 (compound **1c**) and 997537 (compound **1d**), and full synthetic, electrochemical, device theory, device modelling and electronic device data is given in the Supplementary Information. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to L.C. (lee.cronin@glasgow.ac.uk).

LETTER

doi:10.1038/nature13963

Evolution and forcing mechanisms of El Niño over the past 21,000 years

Zhengyu Liu[1,2], Zhengyao Lu[2], Xinyu Wen[2], B. L. Otto-Bliesner[3], A. Timmermann[4] & K. M. Cobb[5]

The El Niño Southern Oscillation (ENSO) is Earth's dominant source of interannual climate variability, but its response to global warming remains highly uncertain[1]. To improve our understanding of ENSO's sensitivity to external climate forcing, it is paramount to determine its past behaviour by using palaeoclimate data and model simulations. Palaeoclimate records show that ENSO has varied considerably since the Last Glacial Maximum (21,000 years ago)[2–9], and some data sets suggest a gradual intensification of ENSO over the past ~6,000 years[2,5,7,8]. Previous attempts to simulate the transient evolution of ENSO have relied on simplified models[10] or snapshot[11–13] experiments. Here we analyse a series of transient Coupled General Circulation Model simulations forced by changes in greenhouse gasses, orbital forcing, the meltwater discharge and the ice-sheet history throughout the past 21,000 years. Consistent with most palaeo-ENSO reconstructions, our model simulates an orbitally induced strengthening of ENSO during the Holocene epoch, which is caused by increasing positive ocean–atmosphere feedbacks. During the early deglaciation, ENSO characteristics change drastically in response to meltwater discharges and the resulting changes in the Atlantic Meridional Overturning Circulation and equatorial annual cycle. Increasing deglacial atmospheric CO_2 concentrations tend to weaken ENSO, whereas retreating glacial ice sheets intensify ENSO. The complex evolution of forcings and ENSO feedbacks and the uncertainties in the reconstruction further highlight the challenge and opportunity for constraining future ENSO responses.

To understand ENSO's evolution during the past 21 kyr, we analyse the baseline transient simulation (TRACE) conducted with the Community Climate System model version 3 (CCSM3). This simulation uses the complete set of realistic climate forcings: orbital, greenhouse gases, continental ice sheets and meltwater discharge (Fig. 1a, d and Methods). TRACE has been shown to replicate many key features of the global climate evolution[14,15]. Over the tropical Pacific, the annual mean sea surface temperature (SST) closely tracks atmospheric CO_2 (Fig. 1d); the cross-equatorial eastern Pacific meridional SST gradient largely tracks the meltwater forcing and the resulting change in the Atlantic Meridional Overturning Circulation (AMOC), a situation consistent with proxy evidence[15] (Fig. 1b, c). ENSO amplitude changes in a complex pattern, as seen in the interannual SST variability over the central-eastern Pacific (Fig. 1e, Extended Data Fig. 1 and Methods). In 100-year windows, ENSO amplitude varies considerably on a multitude of timescales (Extended Data Fig. 2a), in a similar manner to those in other multi-millennial simulations[16] and in palaeo-ENSO reconstructions from lake sediments[2,7,17] (Fig. 1e) and fossil corals[6] from ENSO-teleconnected regions. The unforced variance changes in ENSO may originate from the nonlinear dynamics of ENSO[10] and/or stochastic climate forcings[18,19].

Beyond the background irregularity of ENSO amplitude, however, there are externally forced changes (Fig. 1e and Extended Data Fig. 2a), such as a reduction in ENSO amplitude just after the Last Glacial Maximum (LGM) before levelling off during the Heinrich Stadial 1 (HS1, ~17 kyr ago), a rapid weakening related to the AMOC resumption at the onset of the Bølling–Allerod (BA, ~14.5 kyr ago) and an increase during the Younger Dryas (YD, ~12.9–11.7 kyr ago). During the Holocene,

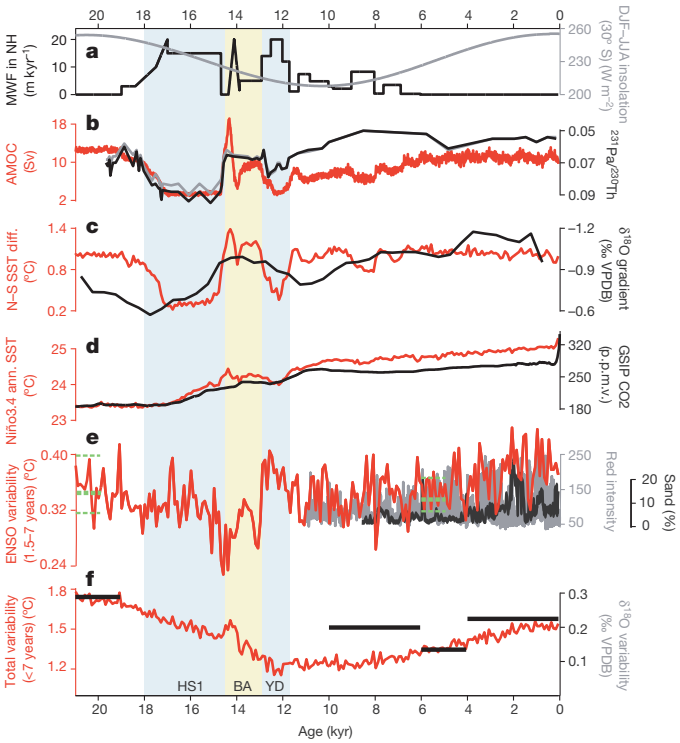


Figure 1 | TRACE simulation and observation. **a**, Grey, amplitude of annual cycle of insolation (December, January and February minus June, July and August (DJF–JJA)) at 30° S (Methods); black, meltwater flux into the North Atlantic (in equivalent sea level change in metres per 1,000 years). **b**, Red, AMOC transport (in sverdrups (Sv); 1 Sv = 10^6 m^3 s^{-1}); black and grey, $^{231}Pa/^{230}Th$ ratio in Bermuda (Methods). **c**, Red, eastern Pacific north–south difference in annual SST (5–15° N minus 5°–15° S, 140–100° W); black, reconstruction[5]. VPDB, Vienna Pee Dee Belemnite. **d**, Red, Niño3.4 (170–120° W, 5° S–5° N) annual SST; black, Niño3.4 CO_2 reconstruction (Methods). **e**, Red, ENSO amplitude (standard deviation of Niño3.4 interannual (1.5–7 years) SST variability) in 100-year windows; black, lake sediment records in the eastern Pacific (El Juno)[7]; grey, lake sediment records on the South American coast (Lake Laguna Pallcacocha)[2]. The green bars at 6 and 21 kyr ago represent the median (solid) and 75th and 25th centiles (dashes) of ENSO amplitude changes in the mid-Holocene and LGM experiments in the PMIP2/PMIP3 ensemble[20] (rescaled using TRACE ENSO in the late Holocene 2–0 kyr ago). **f**, Amplitude (standard deviation) of Niño3 (150–90° W, 5° S–5° N) SST total variability (less than 7 years) in 100-year windows. The black bars show the reconstruction of total SST variance derived from sediment cores in the eastern equatorial Pacific[5].

[1]Department of Atmospheric and Oceanic Sciences and Nelson Center for Climatic Research, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA. [2]Laboratory of Climate, Ocean and Atmosphere Studies, School of Physics, Peking University, Beijing, 100871, China. [3]Climate and Global Dynamics Division, National Center for Atmospheric Research, Boulder, Colorado 80307-3000, USA. [4]International Pacific Research Center and Department of Oceanography, School of Ocean and Earth Science and Technology, University of Hawaii at Manoa, Honolulu, Hawaii 96822, USA. [5]School of Earth and Atmospheric Sciences, Georgia Institute of Technology, Atlanta, Georgia 30332, USA.

ENSO gradually intensified by ~15%. The simulated ENSO evolution is qualitatively similar to that in other models at 6 and 21 kyr ago, falling within the spread of the Palaeoclimate Modelling Intercomparison Project 2 (PMIP2)/PMIP3 ensemble[20] (Fig. 1e).

The simulated ENSO intensification over the Holocene is qualitatively consistent with several key ENSO-sensitive proxy records (Fig. 1e), such as the increase in precipitation variability along the South American coast throughout the Holocene[2,4,7] (Extended Data Figs 3a, b, 4c, d and 5c and Methods), and the increase in ENSO variance in the western Pacific fossil corals from the mid to late Holocene[3]. An intensification of ENSO during the Holocene is also inferred from the ensemble spread of SST from foraminifera in the eastern Pacific[5], although this ensemble spread probably more reflects the total SST variability, which is dominated by the annual cycle rather than ENSO there (Fig. 1f and Extended Data Fig. 2b). Nevertheless, the most recent reconstructions of ENSO's evolution over the Holocene suggest that ENSO variance reached a minimum around the mid Holocene[5,6,21] (Fig. 1f and Extended Data Fig. 6b), with a decrease of ~30–50% relative to the late Holocene. Such a trajectory is not inconsistent with previously published palaeo-ENSO data sets, given their small sample size relative to the high level of intrinsic variability in ENSO amplitudes (Extended Data Figs 2b, c and 6 and Methods).

The simulated trend in Niño3.4 variance over the Holocene amounts to +15%, which is consistent with most mid-Holocene PMIP2/PMIP3 experiments, which show increases of 10–15% in ENSO variance from 6 kyr ago to pre-industrial conditions[20] (Fig. 1e). If the response of ENSO to precessional forcing was indeed relatively modest as in current models, the available palaeo-ENSO data sets are as yet too sparse to detect such a subtle shift in the Holocene, given the high level of background variability[6] (Methods). If, in contrast, ENSO did intensify by 50% and did reach the minimum in the mid-Holocene, the inconsistent simulations in current models would imply model deficiencies. The ultimate resolution of the detailed evolution characteristics of ENSO variance in the Holocene will require much more high-resolution palaeoclimate data from ENSO centres of action.

The comparison with observations before the Holocene poses an even deeper challenge because proxy records of ENSO are rare and less consistent. The few coral records scattered across the last glacial period suggest an active ENSO presence[3,22], but the coral records are too short and sparse to provide a robust estimation of ENSO intensity changes[6] (Methods). Other records seem to show conflicting results. The lithic flux rate associated with flood events in a sediment record off Peru has been used to argue for the change in ENSO-related flooding event, which is weaker at the LGM than the late Holocene[8], as also implied by a varve record in east Africa[17]. The lithic flux data further imply that ENSO intensified rapidly from 17 to 13 kyr ago before gradually peaking at 8 kyr ago[8]. However, the variability of eastern equatorial Pacific subsurface temperatures reconstructed from subsurface dwelling planktonic foraminifera suggests a modestly stronger ENSO at the LGM, which then intensifies to peak at 15 kyr ago and then weakens towards the minimum at 8 kyr ago[9]. The PMIP2/PMIP3 experiments also show a wide spread of ENSO's response at the LGM with no statistically significant change in the ensemble mean ENSO at the LGM relative to the late Holocene[20]. The ENSO amplitude in TRACE is reduced by ~0.2 °C at the LGM relative to the late Holocene, well within the spread of the PMIP2/PMIP3 experiments (Fig. 1e). The inconsistent ENSO responses during the deglaciation, among data and among models, could be caused by various factors. The use of palaeoclimate records in remote regions as proxies for ENSO variance should be treated with caution, especially for those sites outside the region of strong ENSO impact (Extended Data Fig. 4a, b and Methods), where the trend of precipitation variance is dominated by the local response to deglacial climate forcings (Extended Data Fig. 5a, b), rather than by ENSO effects. In addition, the relation between the proxy variance and ENSO variance could change over time, as seen in the less consistent variance between precipitation and ENSO during the deglaciation than during the Holocene in the model (Extended Data Figs 3c–e and 5a, b and Methods). The intrinsic irregularity of ENSO[16] itself also calls into question the use of snapshots of ENSO amplitude, both in palaeoclimate data and in climate models, to represent the continuous evolution.

To determine the physical mechanisms that cause millennial to orbital-scale changes in ENSO strength in TRACE, we quantify the time evolution of ocean–atmosphere feedbacks in the eastern Pacific by using the Bjerknes (BJ) index, which consists of three positive feedbacks (upwelling feedback, zonal advection feedback and thermocline feedback) and two negative feedbacks (heat flux feedback and mean advection feedback)[23] (Methods). In the Holocene, both the BJ index (Fig. 2d) and ENSO amplitude (Fig. 2b) increase in unison (positive correlation of 0.5; $P = 0.004$, 10–0 kyr ago), suggesting a contribution of positive ocean–atmosphere feedback to ENSO intensification. The BJ index increases primarily as a result of the upwelling feedback (Extended Data Fig. 7b). All the positive feedbacks increase throughout the Holocene by means of increased wind sensitivity to SST (Extended Data Fig. 8a), probably related to the warming trend in annual mean climatology (Fig. 1d) that favours active moisture convection and in turn an enhanced atmospheric response to SST anomalies[11]. The tropical warming is forced mainly by the increased annual mean insolation in response to the decreasing obliquity[24] and increasing CO_2. The upwelling feedback is further amplified by the intensified stratification of the upper ocean (Extended Data Fig. 8f), which is generated by the increased austral winter cooling over the subtropical South Pacific (Fig. 2a) and the subsequent equatorward ventilation[12] (Extended Data Fig. 9 and Methods). Thus, the intensifying ENSO in the Holocene is caused mainly by enhanced positive ocean–atmosphere feedbacks, especially the upwelling feedback, in response to precessional forcing.

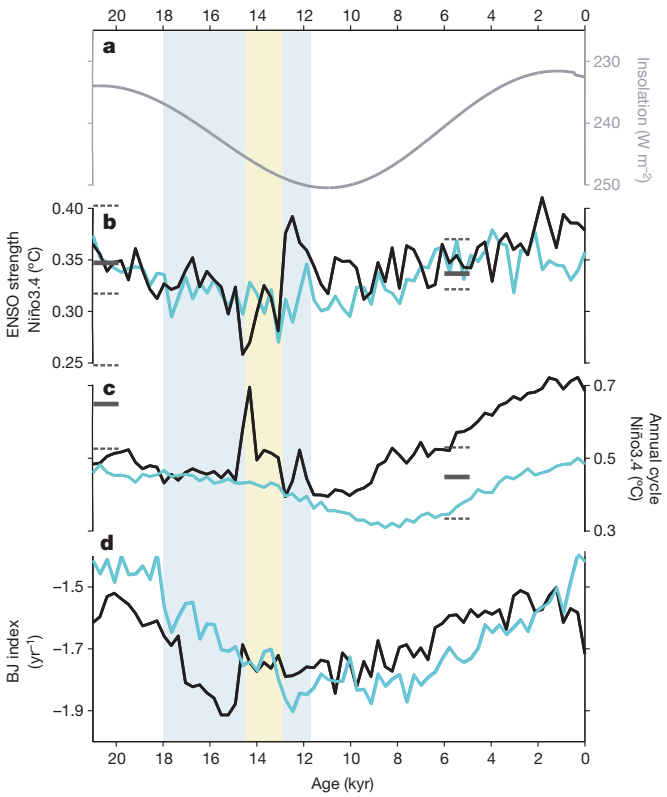


Figure 2 | ENSO, BJ index and annual cycle. **a**, Austral winter (JJA) insolation at 30° S. **b**–**d**, The amplitude of ENSO (**b**), the annual cycle of Niño3.4 SST (**c**) and the BJ index (**d**) in the eastern equatorial Pacific (180–80° W, 5° S–5° N). All are calculated in 300-year windows with black for TRACE and cyan for ORB. In **b** and **c** the grey bars at 6 and 21 kyr ago represent the median (solid) and 75th and 25th centiles (dashes) of the amplitudes of ENSO and annual cycle in the mid-Holocene and LGM experiments in the PMIP2/PMIP3 ensemble[20] (rescaled using TRACE in the late Holocene 2–0 kyr ago).

The dominant role of precessional forcing on ENSO evolution is further confirmed by a sensitivity experiment forced by only orbital forcing (ORB; Methods). The ORB experiment almost reproduces the slow trends of ENSO (Fig. 2b), the annual cycle (Fig. 2c) and the BJ index (Fig. 2d) in TRACE throughout the 21 kyr, except for the millennial-scale events associated with changes in the AMOC during the early deglaciation (Fig. 2b, c). Therefore, the overall evolution of ENSO, with an initial weakening towards the early Holocene and a subsequent strengthening towards the late Holocene, is determined largely by the strength of the ocean–atmosphere feedback (Fig. 2d) and its response to precessional forcing (Fig. 2a). The annual cycle of eastern Pacific SST also weakens towards the early Holocene and recovers towards the late Holocene. This change in annual cycle is caused by the annual cycle of insolation in the subtropical South Pacific (Fig. 1a), which forces a local annual cycle that propagates equatorwards as a result of air–sea interactions[25]. The minimum amplitude of annual cycle in the early to mid Holocene is also qualitatively consistent with PMIP2/PMIP3 experiments for 6 kyr ago and the LGM (Fig. 2c).

In contrast to ENSO's response to precessional forcing, the millennial variability of ENSO's intensity during the early deglaciation is caused by the changes in the annual cycle amplitude triggered by deglacial meltwater fluxes and the resulting AMOC responses. Indeed, the millennial swings of ENSO amplitude often change out of phase with the BJ index, notably around HS1 and YD (Fig. 2b, d and Extended Data Fig. 7a), and therefore cannot be explained by changing ocean–atmosphere instability as with orbital forcing. There is thus no significant correlation between ENSO intensity and the BJ index (0.09; $P = 0.58$, 21–10 kyr ago). Instead, the millennial ENSO variance (Fig. 2b) tends to vary out of phase with the amplitude of the annual cycle (Fig. 2c), with a highly significant negative correlation of −0.49 ($P = 0.002$, 21–10 kyr ago). The close interaction between ENSO and the annual cycle is also consistent with the strong phase-locking of ENSO during early deglaciation (Extended Data Fig. 10). A meltwater pulse in the North Atlantic (Fig. 1a) decreases the AMOC (Fig. 1b), shifts the Intertropical Convergence Zone (ITCZ) southwards and creates an equatorially more symmetric annual mean climate in the eastern equatorial Pacific, corresponding to a weaker north–south difference in SST (Fig. 1c) and a weaker equatorial annual cycle (Fig. 2c)[26]; the weaker annual cycle then amplifies ENSO through the nonlinear mechanism of frequency entrainment[26,27]. The anti-correlation between annual cycle strength and ENSO variance has also been observed for other external forcings in many Coupled General Circulation Models (CGCMs)[28]. The role of the meltwater forcing is confirmed explicitly in a transient sensitivity experiment forced by the deglacial meltwater flux alone (MWF). In the MWF, ENSO's amplitude (Fig. 3a, blue) varies largely out of phase with that of the annual cycle (Fig. 3b, blue) at millennial timescales, in a similar manner to TRACE.

The dominant roles of precessional and meltwater forcings for the deglacial evolution of ENSO raises the question: what is the role of CO_2 forcing on ENSO? To address this question, we analyse an additional sensitivity experiment forced by the greenhouse gases alone (CO2). In CO2, ENSO weakens gradually from 17 to 12 kyr ago (Fig. 3a), following the rising CO_2 (Fig. 1c). The ENSO weakening in CCSM3 can be caused by a more diffusive equatorial thermocline forced by the CO_2 warming[29], and a more asymmetric tropical warming around the Equator and in turn a stronger annual cycle (Fig. 3b) through frequency entrainment[30]. This CO_2-induced ENSO weakening can be detected in TRACE as the gradual reduction of ENSO from 17 kyr ago towards the BA (~14.5 kyr ago). However, the weakening leaves little net signal after the BA (Fig. 3a), because of the offset by the ENSO amplification associated with the retreat of the ice sheet at ~14 kyr ago. As shown in another sensitivity experiment forced by the ice sheet alone (ICE), ENSO is amplified abruptly at 14 kyr ago by a large retreat of the ice sheet over North America (Fig. 3a). The retreat of the ice sheet changes the atmospheric jet and, in turn, the tropical Pacific climatology through teleconnections, which lead to a weaker equatorial annual cycle (Fig. 3b and Methods) and eventually a strengthening ENSO again through frequency entrainment[27]. The

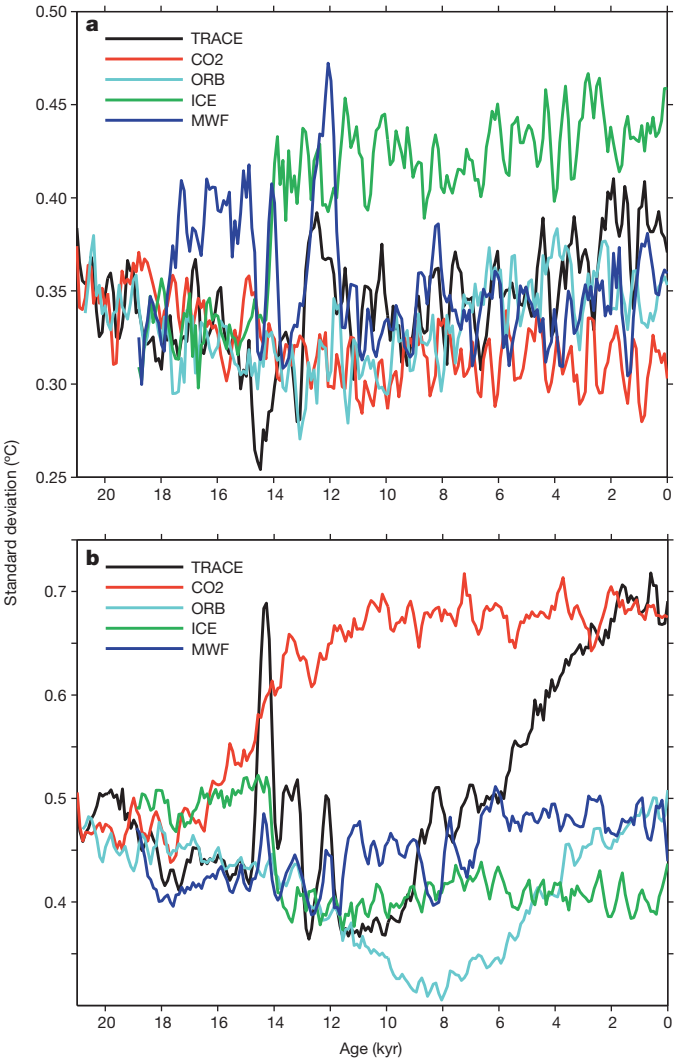


Figure 3 | ENSO in single forcing experiments. **a**, Amplitude of ENSO. **b**, Amplitude of the annual cycle. The amplitude is calculated as the standard deviation of Niño3.4 SST in the frequency band of 1.5–7 years for ENSO, and for the composite seasonal cycle for the annual cycle, both in 300-year windows. The natural variability of the ENSO amplitude can be estimated approximately from the Holocene part of CO2, MWF and ICE as a standard deviation of ~0.2 °C.

opposite effects of CO_2 and the ice sheet on ENSO amplitude offer an explanation of why, unlike the robust ENSO response in the Holocene, the response of ENSO amplitude at the LGM differs between models[20]. At the LGM, because the precessional forcing is similar to that in the late Holocene, ENSO is affected by two large forcings with opposite effects: the lower CO_2 and the presence of large ice sheets.

Overall, the ENSO evolution reconstructed from available proxy records seems to be qualitatively consistent with the simulated ENSO strengthening during the Holocene. However, the proxy data must be improved significantly to help constrain climate models better for the simulation of ENSO evolution in the past and, eventually, for the model projections of the future.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 30 June; accepted 7 October 2014.

1. Collins, M. *et al.* The impact of global warming on the tropical Pacific ocean and El Nino. *Nature Geosci.* **3,** 391–397 (2010).
2. Moy, C. *et al.* Variability of El Niño/Southern Oscillation activity at millennial timescales during the Holocene epoch. *Nature* **420,** 162–166 (2002).

3. Tudhope, A. *et al.* Variability in the El Niño–Southern Oscillation through a glacial–interglacial cycle. *Science* **291,** 1511–1517 (2001).
4. Riedinger, M. *et al.* A ~6100 ^{14}C yr record of El Niño activity from the Galapagos islands. *J. Paleolimnol.* **27,** 1–7 (2002).
5. Koutavas, A. & Joanides, S. El Niño–Southern Oscillation extrema in the Holocene and Last Glacial Maximum. *Paleoceanography* **27,** PA4208, http://dx.doi.org/10.1029/2012PA002378 (2012).
6. Cobb, K. *et al.* Highly variable El Niño–Southern Oscillation throughout the Holocene. *Science* **339,** 67–70 (2013).
7. Conroy, J. *et al.* Holocene changes in eastern tropical Pacific climate inferred from a Galapagos lake sediment record. *Quat. Sci. Rev.* **27,** 1166–1180 (2008).
8. Rein, B. *et al.* El Niño variability off Peru during the last 20,000 years. *Paleoceanography* **20,** PA4003, http://dx.doi.org/10.1029/2004PA001099 (2005).
9. Sadekov, A. *et al.* Paleoclimate reconstructions reveal a strong link between El Niño-Southern Oscillation and tropical Pacific mean state. *Nature Commun.* **4,** 2692, http://dx.doi.org/10.1038/ncomms3692 (2013).
10. Clement, A., Seager, R. & Cane, M. Suppression of El Niño during the Mid-Holocene by changes in the Earth's orbit. *Paleoceanography* **15,** 731–737 (2000).
11. Roberts, W. *An Investigation into the Causes for the Reduction in the Variability of the El Niño-Southern Oscillation in the Early Holocene in a Global Climate Model.* PhD thesis, Univ. Washington (2007).
12. Liu, Z., Kutzbach, J. & Wu, L. Modeling climatic shift of El Niño variability in the Holocene. *Geophys. Res. Lett.* **27,** 2265–2268 (2000).
13. Otto-Bliesner, B. *et al.* Modeling El Niño and its tropical teleconnections during the glacial–interglacial cycle. *Geophys. Res. Lett.* **30,** 10.1029/2003GL08553 (2003).
14. Liu, Z. *et al.* Transient simulation of last deglaciation with a new mechanism for Bølling–Allerød warming. *Science* **325,** 310–314 (2009).
15. Shakun, J. *et al.* Global warming preceded by increasing CO_2 during the last deglaciation. *Nature* **484,** 49–54 (2012).
16. Wittenberg, A. Are historical records sufficient to constrain ENSO simulations. *Geophys. Res. Lett.* **36,** L12702 (2009).
17. Wolff, C. *et al.* Reduced interannual rainfall variability in East Africa during the Last Ice Age. *Science* **333,** 743–747 (2011).
18. Penland, C. & Sardeshmukh, P. The optimal growth of tropical sea surface temperature anomalies. *J. Clim.* **8,** 1999–2024 (1995).
19. Chiang, J., Fang, Y. & Chang, P. Pacific climate change and ENSO activity in the mid-Holocene. *J. Clim.* **22,** 923–939 (2009).
20. Masson-Delmotte, V. *et al.* In *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Stocker, T. F. *et al.*) 383–464 (Cambridge Univ. Press, 2013).
21. Carre, M. *et al.* Holocene history of ENSO variance and asymmetry in the eastern tropical Pacific. *Science* **345,** 1045–1048 (2014).
22. Felis, T. *et al.* Pronounced interannual variability in tropical South Pacific temperatures during Heinrich Stadial 1. *Nature Commun.* **3,** 965, http://dx.doi.org/10.1038/ncomms1973 (2012).
23. Kim, S. & Jin, F. An ENSO stability analysis. Part I: results from a hybrid coupled model. *Clim. Dyn.* **36,** 1593–1607 (2011).
24. Liu, Z., Brady, E. & Lynch-Steiglitz, J. Global ocean response to orbital forcing in the Holocene. *Paleoceanography* **18,** 1041 (2003).
25. Liu, Z. & Xie, S. Equatorward propagation of coupled air–sea disturbances with application to the annual cycle of the eastern tropical Pacific. *J. Atmos. Sci.* **51,** 3807–3822 (1994).
26. Timmermann, A. *et al.* The influence of a weakening of the Atlantic Meridional Overturning Circulation on ENSO. *J. Clim.* **20,** 4899–4919 (2007).
27. Liu, Z. A simple model study of the forced response of ENSO to an external periodic forcing. *J. Clim.* **15,** 1088–1098 (2002).
28. Timmermann, A. *et al.* The effect of orbital forcing on the mean climate and variability of the tropical Pacific. *J. Clim.* **20,** 4147–4159 (2007).
29. Meehl, G., Teng, H. & Branstator, G. Future changes of El Niño in two global climate models. *Clim. Dyn.* **26,** 549–566 (2006).
30. Timmermann, A., Jin, F. & Collins, M. Intensification of the annual cycle in the tropical Pacific due to greenhouse warming. *Geophys. Res. Lett.* **31,** L12208 (2004).

Acknowledgements This work is supported by the US National Science Foundation (NSF)/P2C2 Program, Chinese NSFC41130105, the US Department of Energy/Office of Science (BER), Chinese MOST2012CB955200, NSF1049219 and 1204011. The computation is carried out at Oak Ridge National Laboratory of the Department of Energy and the National Center for Atmospheric Research supercomputing facility.

Author Contributions Z. Liu conceived the study and wrote the paper. Z. Lu and XW performed the analysis. Z. Liu and B.O.B. contributed to the simulations. A.T. and K.C. contributed to the interpretation. All authors discussed the results and provided inputs to the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Z. Liu (zliu3@wisc.edu).

LETTER

doi:10.1038/nature13761

Multiplex single-molecule interaction profiling of DNA-barcoded proteins

Liangcai Gu[1], Chao Li[2], John Aach[1], David E. Hill[1,3], Marc Vidal[1,3] & George M. Church[1,2]

In contrast with advances in massively parallel DNA sequencing[1], high-throughput protein analyses[2–4] are often limited by ensemble measurements, individual analyte purification and hence compromised quality and cost-effectiveness. Single-molecule protein detection using optical methods[5] is limited by the number of spectrally non-overlapping chromophores. Here we introduce a single-molecular-interaction sequencing (SMI-seq) technology for parallel protein interaction profiling leveraging single-molecule advantages. DNA barcodes are attached to proteins collectively via ribosome display[6] or individually via enzymatic conjugation. Barcoded proteins are assayed en masse in aqueous solution and subsequently immobilized in a polyacrylamide thin film to construct a random single-molecule array, where barcoding DNAs are amplified into *in situ* polymerase colonies (polonies)[7] and analysed by DNA sequencing. This method allows precise quantification of various proteins with a theoretical maximum array density of over one million polonies per square millimetre. Furthermore, protein interactions can be measured on the basis of the statistics of colocalized polonies arising from barcoding DNAs of interacting proteins. Two demanding applications, G-protein coupled receptor and antibody-binding profiling, are demonstrated. SMI-seq enables 'library versus library' screening in a one-pot assay, simultaneously interrogating molecular binding affinity and specificity.

To analyse proteins in a massively parallel single-molecule format, we generated proteins that are molecularly coupled to a DNA bearing a barcoding sequence. One barcoding approach is to translate and display proteins on protein–ribosome–messenger-RNA–complementary-DNA (PRMC) complexes *in vitro*, in which the cDNA contains a synthetic barcode at the 5′ end of protein open reading frames (ORFs) (Fig. 1a). Specifically, the ribosome display was performed by using mRNA–cDNA hybrids as templates and an *in vitro* translation (IVT) system reconstituted with purified components[8] that was shown to stabilize PRMC complexes (Extended Data Fig. 1). PRMC complexes bearing full-length proteins of interest were enriched by Flag-tag affinity purification. Notably, this approach is applicable to a library of proteins of various sizes and size-related biases during decoding can be avoided by using uniformly sized barcoding DNAs. Alternatively, some proteins that can only be functionally expressed *in vivo* require individual barcoding. Thus, fusion proteins were constructed with an engineered enzyme tag, HaloTag[9], which mediates an efficient covalent conjugation to a HaloTag-ligand-modified double-stranded DNA (Fig. 1b). Our method is adaptable to a microtitre plate format for automated parallel protein production (Extended Data Fig. 2).

A complex mixture of barcoded proteins can be identified and quantified by *in situ* sequencing of their barcodes (Fig. 2a). The proteins were immobilized into an ultrathin layer of crosslinked polyacrylamide gel attached to a microscopic slide, and their barcoding DNAs bearing a 5′-acrydite modification (Fig. 1) were covalently crosslinked to the gel matrix to prevent template drifting (Extended Data Fig. 3). A solid-phase PCR, with two gel-anchored primers, was performed according to an adapted isothermal bridge amplification protocol[10] in an assembled flow cell. This amplification showed a high efficiency of ~80% barcode detection (Extended Data Fig. 4a), and resulted in polonies of ~1 μm diameter (Fig. 2b), similar to the clusters generated on an Illumina platform[10]. Polonies were identified by hybridization with fluorescent probes, single-base extension (SBE) or ligation-based sequencing[11].

To test the accuracy of our method, we selected nine immunoglobulin and non-immunoglobulin binding proteins and three antigens (for example, human, bacterial and viral proteins) of a molecular weight ranging from 3.4 to 120 kDa (Extended Data Table 1). Mixed PRMC complexes were prepared in six barcoded dilutions, with concentrations spanning six orders of magnitude, pooled together and subjected to the single-molecule quantification. Barcode detection efficiencies of different proteins were found to be almost identical at various concentrations (Extended Data Fig. 4). The *in situ* single-molecule quantification can avoid PCR amplification bias[12] and shows high reproducibility; the Pearson correlation coefficient r was above 0.99 when over 1,000 protein polonies were detected (Fig. 2c). Because proteins were highly diluted (at less than picomolar concentrations) before array deposition, protein monomers should be the predominant form.

Interacting barcoded proteins can be indirectly detected by joining their barcoding DNAs via ligation[13,14] or primer extension[15]. Here, direct observation and counting of single-molecule protein complexes should

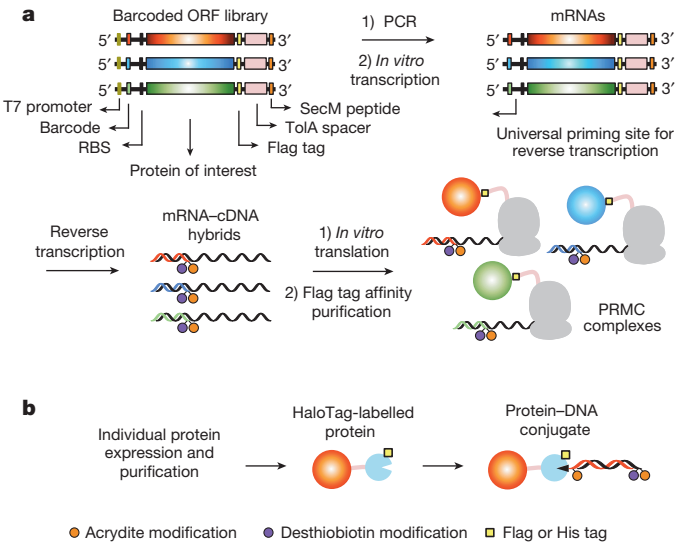


Figure 1 | Schematics of protein barcoding methods. a, Collective barcoding via ribosome display. A short synthetic barcoding sequence is joined to the 5′ end of DNA templates via PCR. PRMC complexes are formed via ribosome stalling triggered by a carboxy-terminal *Escherichia coli* SecM peptide. Displayed proteins bearing a C-terminal Flag tag are separated from the ribosomes by an *E. coli* TolA spacer domain. RBS, ribosomal binding site. **b**, Individual barcoding via a HaloTag-mediated conjugation of proteins to a 220-base-pair (bp) double-stranded barcoding DNA with a HaloTag ligand modification (black triangle). Modifications are introduced to barcoding DNAs by PCR with modified primers.

[1]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [2]Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, Massachusetts 02115, USA. [3]Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA.

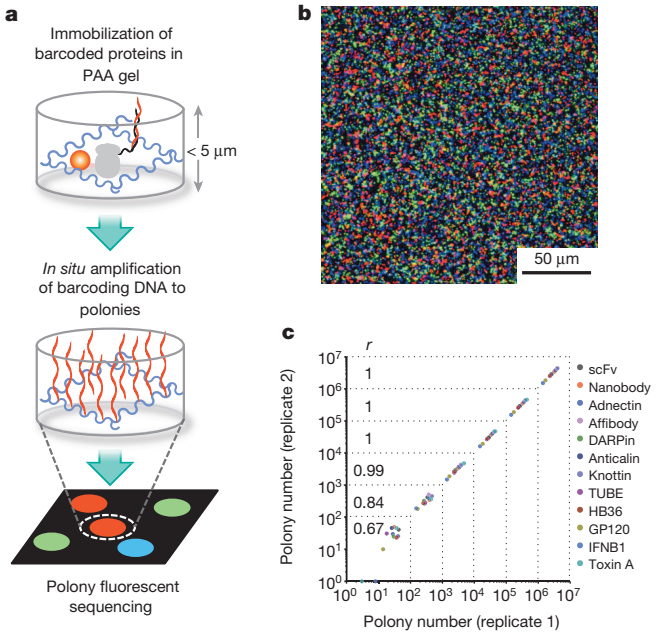


Figure 2 | Amplification and quantification of barcoding DNAs. **a**, Schematic of *in situ* polony amplification and sequencing. Barcoded proteins were immobilized in a polyacrylamide (PAA) gel matrix attached to a Bind-Silane-treated glass slide. The slide was assembled into a flow cell, where barcoding DNAs were amplified *in situ* into polonies for DNA sequencing. **b**, Representative merged images of polonies hybridized with Cy5 (red), Cy3 (green) and fluorescein (blue)-labelled oligonucleotides (×20 objective magnification). **c**, Polony quantification of mixed protein binders and antigens. The Pearson correlation coefficient *r* was calculated for different coverages grouped by dotted lines.

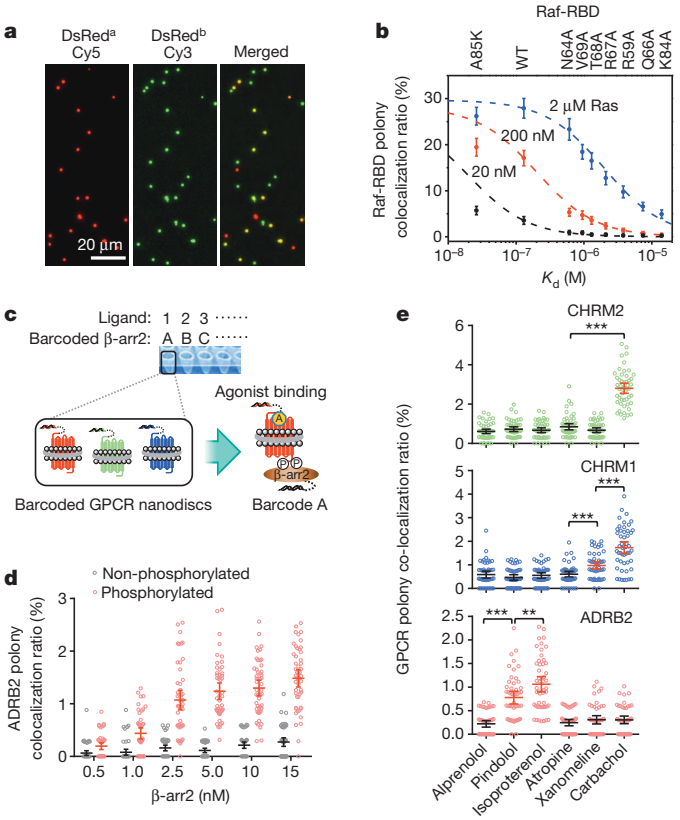


Figure 3 | Analyses of protein interactions via polony colocalization. **a**, Interaction of DsRed subunits resulted in colocalized polonies. Polonies of the differently barcoded subunits, DsRed[a] and DsRed[b], were identified by SBE with, respectively, Cy5 (red) or Cy3 (green)-labelled dideoxynucleotide triphosphates (ddNTPs). **b**, Correlation between the polony colocalization ratios and K_d values of Ras–Raf-RBD complexes. Means of measurements at 100 imaging positions ± 95% confidence level (CL; refer to Supplementary Table 1). Fitting equation, $R = R_{max} \times P/(K_d + P)$, where R is the predicted Raf-RBD polony colocalization ratio, R_{max} is the maximum polony colocalization ratio when Raf-RBD is saturated by Ras, and P is the Ras concentration. WT, wild type. **c**, Schematic of multiplex GPCR screening and compound profiling by the binding assay of mixed barcoded GPCRs with barcoded β-arr2. **d**, Comparison of β-arr2 binding to isoproterenol-activated β_2-adrenergic receptor with or without GRK2-mediated phosphorylation. Titration data of β-arr2 were fitted by the one-site-specific model using GraphPad Prism 6. **e**, Parallel GPCR binding profiling. Data represent mean values of 50 measurements; error bars, 95% CL, highlighted in red for agonists (refer to Supplementary Table 2). $^{**}P < 0.01$, $^{***}P < 0.001$, one-tailed paired Student's *t*-test.

be possible if barcoding DNAs of interacting proteins can be amplified into colocalized polonies. To test this, we generated DsRed, which naturally forms a tetramer, with monomers each bearing one of two different barcodes. To avoid dissociation of any complexes during the array analysis, we crosslinked them with an amine-reactive crosslinker, bis-*N*-succinimidyl-(pentaethylene glycol) ester ($BS(PEG)_5$). The crosslinking was shown to be efficient due to the presence of a lysine-rich TolA spacer domain (Fig. 1a and Extended Data Fig. 5). It is evident that barcoding DNAs of the colocalized monomers (DsRed[a] and DsRed[b]) were co-amplified into overlapping polonies (Fig. 3a), providing a solid basis for further applications.

Unlike other methods that only detect affinity-enriched proteins (for example, PLATO[16]), our approach simultaneously counts polonies of both unbound and bound proteins in a single solution. Thus, we sought to determine if it can provide a measure of protein binding affinities. We chose a model system, the GTP-dependent binding of human H-Ras (Ras) to Ras-binding domain of c-Raf-1 (Raf-RBD)[17]. A Raf-RBD polony colocalization ratio—the percentage of Raf-RBD polonies colocalized with Ras polonies—was measured for wild-type Ras and Raf-RBD and eight Raf-RBD mutants; the Ras protein concentration was titrated over three orders of magnitude (Fig. 3b). Although the colocalization ratio is dependent on protein concentration and crosslinking efficiency and can be affected by experimental variables (protein quality, crosslinking conditions, polony array density, etc.), all the proteins within a single assay are under the same reaction conditions. Given a similar proportion of active protein and crosslinking efficiency, polony colocalization ratios could be correlated with ratios of bound proteins at equilibrium and thus their binding affinities. To test this, the colocalization ratios were plotted against previously reported dissociation constants (K_d values) ranging from nanomolar to micromolar values[18,19] (Fig. 3b and Supplementary Table 1), and fitted by using a one-site-specific binding model (dashed curves). The fitted and observed average colocalization ratios show relatively high agreement ($r > 0.96$), except for the A85K mutant which displayed significantly lower experimental values than predicted by the model, probably owing to the disruption of Lys 85-mediated interactions[19] by the crosslinking. Therefore, this method could be useful for high-throughput screening of protein binding affinities.

As a first high-throughput screening application, we investigated small molecule-mediated protein–protein interactions. An advantage of our method over traditional solid-phase techniques such as protein microarrays[3] is that we store and assay proteins in an aqueous solution. To exploit this, we decided to address the challenges in screening G-protein coupled receptors (GPCRs), the largest membrane protein family and premier drug targets[20]. Current GPCR–ligand screening techniques mainly rely on cell-based assays[21], which are subject to limitations such as the heterogeneous nature of the samples, the presence of other cellular components that can cause false positives or negatives, and limited miniaturization and multiplexing capability (for example, one receptor per

assay). To prepare a homogeneous single-molecule GPCR sample compatible with our approach, receptors were stabilized in phospholipid bilayer nanodiscs[22] by assembling detergent-solubilized GPCRs, phospholipids and a membrane scaffold protein, MSP1E3D1, into GPCR–nanodisc complexes[23,24]. GPCR activation upon ligand binding can be functionally assessed by β-arrestin binding to activated receptors, which is a G-protein-independent assay applicable to almost all GPCRs, including orphan receptors[25].

A compound library can be screened in multi-well plates, and in each well one compound is assayed with many barcoded GPCRs and a β-arrestin-2 (β-arr2) protein bearing a well-position-associated barcode (Fig. 3c). All the samples were pooled and deposited on one slide, and GPCR agonists were detected by measuring GPCR polony colocalization with corresponding β-arr2 polonies. Our efforts to obtain functional GPCRs using IVT systems were not successful, so they were expressed in baculovirus-infected insect cells, purified in nanodiscs and individually barcoded (Fig. 1b). To establish assay conditions, we examined β-arr2 binding to an agonist (isoproterenol)-saturated β_2-adrenergic receptor (ADRB2), with and without GPCR kinase 2 (GRK2)-mediated receptor phosphorylation and under varied β-arr2 protein concentrations (Fig. 3d). The colocalization ratios were measured at 50 imaging positions on the array for statistical analysis. As expected, coupling the receptor phosphorylation to the assay improves the β-arr2 binding; 3- to 11-fold increases (largest $P = 0.002$) of the average colocalization ratios after phosphorylation were observed. The fitting of β-arr2 titration data for the phosphorylated receptor yielded an apparent K_d of 0.95 nM, which is close to the K_d of 0.23 nM obtained from traditional binding assays using radiolabelled β-arr2[26].

To test the screening performance, we assayed three GPCRs, ADRB2, M1 and M2 muscarinic acetylcholine receptors (CHRM1 and CHRM2), with six compounds including full, partial, subtype-selective and non-selective agonists and two antagonists (Fig. 3e and Supplementary Table 2). The colocalization statistical analysis based on measurements of ~13,000–17,000 polonies for each receptor precisely identified the full agonists (isoproterenol and carbachol) from the others (largest $P < 2.7 \times 10^{-10}$). Moreover, different types of agonists can be distinguished by comparing their polony colocalization ratios, for example, the full and partial agonists of ADRB2 (isoproterenol and pindolol, respectively; $P < 0.004$), and the orthosteric and allosteric agonists of CHRM1 (carbachol and xanomeline, respectively; $P < 3 \times 10^{-6}$). Thus, our method could allow parallel GPCR screening and compound profiling.

An intriguing feature of this approach is the ability to screen two barcoded libraries in a single binding assay. Established techniques (for example, yeast two-hybrid systems[2]) for library versus library screening are cell-based and require pairing both genes from two libraries to identify positive clones by performing individual PCR reactions[27]. To demonstrate this capability, we prototyped a test of a demanding application, the binding profiling of an antibody repertoire. The screening of natural or semisynthetic monoclonal antibody libraries essentially includes binding affinity selection and specificity profiling, which have to be conducted separately with current techniques. The traditional specificity profiling is costly, usually requiring at least one protein chip for a single antibody test[28], and thus has only been commercially applied to therapeutic antibodies. However, both processes could be integrated on our platform by screening an antibody library with a target-protein library.

Specifically, we performed a one-pot assay containing 200 ribosome-displayed single-chain variable fragments (scFvs) and 55 human proteins including cytokines, growth factors and receptors synthesized *in vitro* (Extended Data Table 2). Twenty scFvs were derived by random mutagenesis from each of ten scFvs, the genes of which were previously synthesized from a programmable DNA microchip[29]. We sequenced ~0.64 million polonies and measured the colocalization ratios for 11,000 scFv–target protein (probe) pairs at 100 imaging positions (Fig. 4a and Supplementary Table 3). Of 200 scFvs, 147 were found with the highest colocalization ratios, 95 of which are significantly above the second highest ($P < 0.05$), and thus the highest specificity, to their predicted

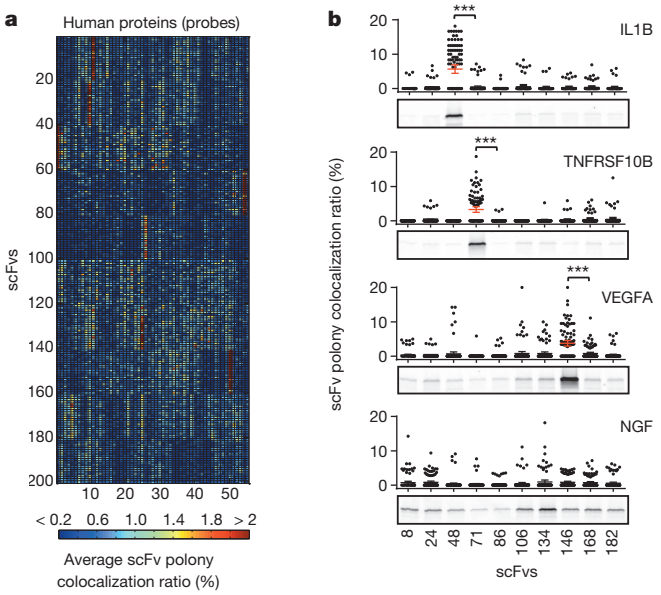


Figure 4 | Parallel antibody binding profiling. **a**, Heat map of the mean colocalization ratios measured at 100 imaging positions (refer to Supplementary Table 3). ScFvs sharing the same origins were grouped by their numbers (Extended Data Table 2). **b**, Correlation between the polony colocalization statistics and the scFv immunoprecipitation results. For the immunoprecipitation assay, selected scFvs were fused to a C-terminal streptavidin binding peptide tag and bound to streptavidin magnetic beads to pull down human protein probes bearing a HaloTag, which can be labelled by Halo- tetramethylrhodamine (TMR) for fluorescent gel imaging. Error bars, 95% CL, highlighted in red for specific scFv–probe binding. ***$P < 0.001$, one-tailed paired Student's *t*-test.

targets; the others failed probably because the construction of scFv fragments and mutations inhibit target binding. Substantial cross-reactivity can be sensitively detected, for example, 3,474 scFv–probe pairs showed tenfold higher polony colocalization than random distribution ($P < 0.01$). scFv mutants of a same scFv, grouped by their numbers, exhibit similar but not identical binding patterns to the probes. Next, we confirmed the results of 40 scFv–probe pairs by individual immunoprecipitation assays and the colocalization statistics were consistent with relative fluorescence intensities of the probe protein bands (Fig. 4b). Moreover, to further assess multiplexing potential, we developed a mathematical model that integrated parameters including K_d values of protein–probe complexes to be detected and numbers of proteins and probes that can be assayed simultaneously (Supplementary Notes). The model suggests that tens of thousands of proteins and probes can be quantifiably analysed within a single assay.

The protein barcoding requirement imposes limitations on SMI-seq. First, it cannot directly analyse proteins from biological samples. However, non-barcoded proteins can be detected in a similar fashion by using barcoded antibodies or aptamers as part of a proximity ligation assay[13,14]. In addition, PRMC complexes are susceptible to nuclease contamination, thus limiting the choice of IVT systems. Finally, barcoding DNA can non-specifically bind to proteins bearing nucleic-acid-binding domains. Although in the present study DNA templates were individually barcoded, a large library can be prepared by introducing millions of chip-synthesized[29] or random barcoding sequences to an open reading frame (ORF) library by a single PCR reaction and later matching them to ORF sequences by next-generation sequencing. SMI-seq enables single-molecule counting of proteins and complexes *in situ*, fundamentally improving sensitivity, accuracy and multiplexity (Extended Data Table 3 and Supplementary Discussion), and thus the demonstrated applications are difficult or impossible to perform with other high-throughput techniques (for example, PLATO). It is readily adaptable to industrial next-generation sequencing platforms and translated into

many applications. In addition to natural and recombinant proteins, it will be applicable to *de novo* proteins (for example, with unnatural amino acids or modifications), nucleic acids and barcoded small molecules[30].

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 11 February; accepted 8 August 2014.

Published online 21 September 2014.

1. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nature Biotechnol.* **26,** 1135–1145 (2008).
2. Dreze, M. *et al.* High-quality binary interactome mapping. *Methods Enzymol.* **470,** 281–315 (2010).
3. MacBeath, G. & Schreiber, S. L. Printing proteins as microarrays for high-throughput function determination. *Science* **289,** 1760–1763 (2000).
4. Gavin, A. C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415,** 141–147 (2002).
5. Weiss, S. Fluorescence spectroscopy of single biomolecules. *Science* **283,** 1676–1683 (1999).
6. Hanes, J. & Pluckthun, A. *In vitro* selection and evolution of functional proteins by using ribosome display. *Proc. Natl Acad. Sci. USA* **94,** 4937–4942 (1997).
7. Mitra, R. D. & Church, G. M. *In situ* localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Res.* **27,** e34–e39 (1999).
8. Shimizu, Y. *et al.* Cell-free translation reconstituted with purified components. *Nature Biotechnol.* **19,** 751–755 (2001).
9. Los, G. V. *et al.* HatoTag: A novel protein labeling technology for cell imaging and protein analysis. *ACS Chem. Biol.* **3,** 373–382 (2008).
10. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).
11. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309,** 1728–1732 (2005).
12. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12,** R18 (2011).
13. Fredriksson, S. *et al.* Protein detection using proximity-dependent DNA ligation assays. *Nature Biotechnol.* **20,** 473–477 (2002).
14. Hammond, M., Nong, R. Y., Ericsson, O., Pardali, K. & Landegren, U. Profiling cellular protein complexes by proximity ligation with dual tag microarray readout. *PLoS ONE* **7,** e40405 (2012).
15. Lundberg, M., Eriksson, A., Tran, B., Assarsson, E. & Fredriksson, S. Homogeneous antibody-based proximity extension assays provide sensitive and specific detection of low-abundant proteins in human blood. *Nucleic Acids Res.* **39,** e102 (2011).
16. Zhu, J. *et al.* Protein interaction discovery using parallel analysis of translated ORFs (PLATO). *Nature Biotechnol.* **31,** 331–334 (2013).
17. Vetter, I. R. & Wittinghofer, A. The guanine nucleotide-binding switch in three dimensions. *Science* **294,** 1299–1304 (2001).
18. Block, C., Janknecht, R., Herrmann, C., Nassar, N. & Wittinghofer, A. Quantitative structure-activity analysis correlating Ras/Raf interaction *in vitro* to Raf activation *in vivo*. *Nature Struct. Biol.* **3,** 244–251 (1996).
19. Kiel, C. *et al.* Improved binding of Raf to Ras·GDP is correlated with biological activity. *J. Biol. Chem.* **284,** 31893–31902 (2009).
20. Overington, J. P., Al-Lazikani, B. & Hopkins, A. L. How many drug targets are there? *Nature Rev. Drug Discov.* **5,** 993–996 (2006).
21. Zhang, R. & Xie, X. Tools for GPCR drug discovery. *Acta Pharmacol. Sin.* **33,** 372–384 (2012).
22. Denisov, I. G., Grinkova, Y. V., Lazarides, A. A. & Sligar, S. G. Directed self-assembly of monodisperse phospholipid bilayer nanodiscs with controlled size. *J. Am. Chem. Soc.* **126,** 3477–3487 (2004).
23. Leitz, A. J., Bayburt, T. H., Barnakov, A. N., Springer, B. A. & Sligar, S. G. Functional reconstitution of β_2-adrenergic receptors utilizing self-assembling nanodisc technology. *Biotechniques* **40,** 601–612 (2006).
24. Whorton, M. R. *et al.* A monomeric G protein-coupled receptor isolated in a high-density lipoprotein particle efficiently activates its G protein. *Proc. Natl Acad. Sci. USA* **104,** 7682–7687 (2007).
25. Luttrell, L. M. & Lefkowitz, R. J. The role of β-arrestins in the termination and transduction of G-protein-coupled receptor signals. *J. Cell Sci.* **115,** 455–465 (2002).
26. Gurevich, V. V. *et al.* Arrestin interactions with G-protein-coupled receptors - Direct binding-studies of wild-type and mutant arrestins with rhodopsin, β_2-adrenergic, and m2-muscarinic cholinergic receptors. *J. Biol. Chem.* **270,** 720–731 (1995).
27. Yu, H. *et al.* Next-generation sequencing to generate interactome datasets. *Nature Methods* **8,** 478–480 (2011).
28. Michaud, G. A. *et al.* Analyzing antibody specificity with whole proteome microarrays. *Nature Biotechnol.* **21,** 1509–1512 (2003).
29. Kosuri, S. *et al.* Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nature Biotechnol.* **28,** 1295–1299 (2010).
30. McGregor, L. M., Jain, T. & Liu, D. R. Identification of ligand-target pairs from combined libraries of small molecules and unpurified protein targets in cell lysates. *J. Am. Chem. Soc.* **136,** 3264–3270 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by a grant from the US Department of Energy (DE-FG02-02ER63445) to G.M.C. and a grant from the NIH NHGRI (HG001715) to M.V. and D.E.H. L.G. was supported by a postdoctoral fellowship from the Jane Coffin Childs Fund for Medical Research and a grant from Harvard Origins of Life Initiative. We thank W. Harper, L. Pontano Vaites, S. Elledge and J. Zhu for providing plasmids, F. Vigneault for comments on the manuscript, J. Lai and D. Breslau for assistance on imaging setup, and members of the Church and Vidal groups for constructive discussions.

Author Contributions L.G. and G.M.C. conceived the technique; L.G. and C.L. performed experiments and analysed data; J.A. built the mathematical model and assisted the colocalization analyses; D.E.H. and M.V. assisted the production of barcoded proteins; L.G., J.A. and G.M.C. wrote the manuscript with help from the other authors.

Author Information MATLAB scripts for imaging analyses, colocalization statistics and mathematical modelling can be found at http://arep.med.harvard.edu/SMI-Seq/. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to L.G. (liangcaigu@gmail.com) or G.M.C. (gchurch@genetics.med.harvard.edu).

LETTER

doi:10.1038/nature13904

MPDL3280A (anti-PD-L1) treatment leads to clinical activity in metastatic bladder cancer

Thomas Powles[1], Joseph Paul Eder[2], Gregg D. Fine[3], Fadi S. Braiteh[4], Yohann Loriot[5], Cristina Cruz[6], Joaquim Bellmunt[7], Howard A. Burris[8], Daniel P. Petrylak[2], Siew-leng Teng[3], Xiaodong Shen[3], Zachary Boyd[3], Priti S. Hegde[3], Daniel S. Chen[3] & Nicholas J. Vogelzang[9]

There have been no major advances for the treatment of metastatic urothelial bladder cancer (UBC) in the last 30 years. Chemotherapy is still the standard of care. Patient outcomes, especially for those in whom chemotherapy is not effective or is poorly tolerated, remain poor[1,2]. One hallmark of UBC is the presence of high rates of somatic mutations[3–5]. These alterations may enhance the ability of the host immune system to recognize tumour cells as foreign owing to an increased number of antigens[6]. However, these cancers may also elude immune surveillance and eradication through the expression of programmed death-ligand 1 (PD-L1; also called CD274 or B7-H1) in the tumour microenvironment[7,8]. Therefore, we examined the anti-PD-L1 antibody MPDL3280A, a systemic cancer immunotherapy, for the treatment of metastatic UBC. MPDL3280A is a high-affinity engineered human anti-PD-L1 monoclonal immunoglobulin-G1 antibody that inhibits the interaction of PD-L1 with PD-1 (PDCD1) and B7.1 (CD80)[9]. Because PD-L1 is expressed on activated T cells, MPDL3280A was engineered with a modification in the Fc domain that eliminates antibody-dependent cellular cytotoxicity at clinically relevant doses to prevent the depletion of T cells expressing PD-L1. Here we show that MPDL3280A has noteworthy activity in metastatic UBC. Responses were often rapid, with many occurring at the time of the first response assessment (6 weeks) and nearly all were ongoing at the data cutoff. This phase I expansion study, with an adaptive design that allowed for biomarker-positive enriched cohorts, demonstrated that tumours expressing PD-L1-positive tumour-infiltrating immune cells had particularly high response rates. Moreover, owing to the favourable toxicity profile, including a lack of renal toxicity, patients with UBC, who are often older and have a higher incidence of renal impairment, may be better able to tolerate MPDL3280A versus chemotherapy. These results suggest that MPDL3280A may have an important role in treating UBC—the drug received breakthrough designation status by the US Food and Drug Administration (FDA) in June 2014.

We report on the safety and activity of MPDL3280A in patients with UBC who were enrolled in a UBC expansion cohort of a large phase I trial with an adaptive design. This progressive design has been used previously to investigate immune checkpoint inhibitors in a spectrum of tumours and has resulted in regulatory approval in other settings[10] (see also http://www.specialtypharmajournal.com/medical-news/oncology/5119-japanese-regulators-approve-the-first-pd-1-drug-for-treatment-of-melanoma). This UBC cohort was initially selected by PD-L1 immunohistochemistry (IHC) on tumour-infiltrating immune cells to test the hypothesis that PD-L1-positive patients might specifically respond to MPDL3280A. The cohort was subsequently expanded to include patients regardless of PD-L1 status to determine whether PD-L1-negative patients could also respond. Overall, 205 patients were pre-screened and specimens were centrally analysed for PD-L1 expression. Of the available pre-screen samples, 59% were from resections and 27% were from biopsies. Analysis was permitted on both archived and fresh tissue. The time between tissue collection and starting MPDL3280A treatment is shown in Extended Data Fig. 1. The prevalence of positive PD-L1 expression (IHC score 2 or 3 (2/3)) in tumour-infiltrating immune cells in the pre-screened population was 27% (Fig. 1a, b). Only 4% of pre-screened patients had positive PD-L1 expression in tumour-infiltrating immune cells and in tumour cells.

Patients in the UBC cohort were dosed between 13 March 2013 and 1 January 2014. As of the clinical cutoff date of 1 January 2014, 68 patients with UBC received treatment and were evaluable for safety. Sixty-seven patients were evaluable for efficacy (one patient had less than 6 weeks follow up and therefore no efficacy evaluation). Of the efficacy-evaluable patients, 12 (18%) had tumours scored as PD-L1 IHC 0, 23 (34%) as IHC 1, 20 (30%) as IHC 2, 10 (15%) as IHC 3, and 2 (3%) as unknown based on tumour-infiltrating immune cells (see Methods for precise definitions). One patient had a PD-L1 IHC score of 2 or 3 for both tumour-infiltrating immune cells and tumour cells. Twenty-one patients with PD-L1 IHC 2 or 3 scores were enrolled before the cohort was expanded to include patients regardless of IHC status. Patients were pre-treated with 62 (93%) receiving previous cisplatin- or carboplatin-based chemotherapy (53 (79%) received previous cisplatin) and 48 (72%) receiving 2 or more previous systemic treatments. Furthermore, many patients had poor prognostic factors at baseline[11,12], including 50 (75%) with visceral metastases, 12 (19%) with haemoglobin levels less than 10 g dl^{-1}, 22 (33%) with creatinine clearance less than 60 ml min^{-1}, 39 (59%) with an Eastern Cooperative Oncology Group (ECOG) performance score of 1, and 26 (42%) whose time from previous chemotherapy was 3 months or less (Table 1).

In the safety-evaluable population, patients with UBC received MPDL3280A for a median duration of 65 days (range: 1–259 days). Of these patients, 57% reported a treatment-related adverse event (AE) of any grade, and 4% reported a grade 3 treatment-related AE, which included one occurrence each of asthenia, thrombocytopaenia and decreased blood phosphorus (Table 2 and Extended Data Table 1). There were no grade 4 or 5 treatment-related AEs. Most treatment-related AEs were grade 1 or 2, and many were transient in nature. Overall, decreased appetite (grade 1/2, 22%; grade 3/4, 0%) and fatigue (grade 1/2, 18%; grade 3/4, 0%) were the most commonly reported toxicities and are thought to be related to immune system activation[13] (Extended Data Table 2). No investigator-assessed immune-related toxicities were reported.

For patients with a minimum of 6 weeks of follow-up, objective response rates (ORRs) were 43% (13 of 30; 95% confidence interval (CI): 26–63%) for those with IHC 2/3 tumours and 11% (4 of 35; 95% CI: 4–26%) for those with IHC 0 or 1 (0/1) tumours. The IHC 2/3 ORR included a 7% complete response rate (2 of 30) (Figs 1c and 2). Among

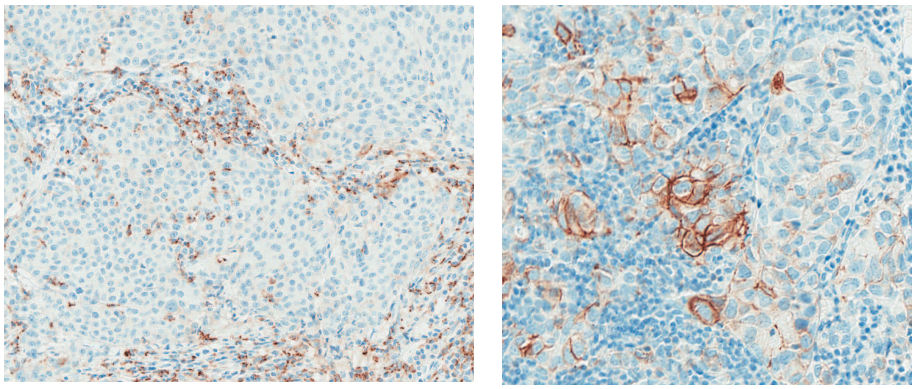
[1]Barts Cancer Institute, Queen Mary University of London, Barts Experimental Cancer Medicine Centre, London EC1M 6BQ, UK. [2]Yale Cancer Center, 333 Cedar Street, WWW211, New Haven, Connecticut 06520, USA. [3]Genentech, Inc. 1 DNA Way, South San Francisco, California 94080, USA. [4]Comprehensive Cancer Centers of Nevada, 3730 S. Eastern Avenue, Las Vegas, Nevada 89169, USA. [5]Gustave Roussy, 114 Rue Édouard Vaillant, 94805 Villejuif, France. [6]Vall d'Hebron Institute of Oncology (VHIO) and Vall d'Hebron University Hospital. Passeig Vall d'Hebron, 119-129, 08035, Barcelona, Spain. [7]Bladder Cancer Center, Dana-Farber/Brigham and Women's Cancer Center, Harvard Medical School, 450 Brookline Avenue, Boston, Massachusetts 02215, USA. [8]Sarah Cannon Research Institute, 3322 West End Avenue, Suite 900, Nashville, Tennessee 37203, USA. [9]University of Nevada School of Medicine and US Oncology/Comprehensive Cancer Centers of Nevada, 3730 S. Eastern Avenue, Las Vegas, Nevada 89169, USA.

a

PD-L1 prevalence in UBC tumours by IHC

n = 205	PD-L1-positive tumour-infiltrating immune cells (no. of specimens (%))	PD-L1-positive tumour cells (no. of specimens (%))
IHC 3	18 (9)	14 (7)
IHC 2	37 (18)	8 (4)
IHC 1	89 (43)	37 (18)
IHC 0	61 (30)	146 (71)

b



Tumour-infiltrating immune cells Tumour cells

c

Tumour-infiltrating immune cells and objective response rates

	Objective response rate *n* (%)	Stable disease *n* (%)	Progressive disease *n* (%)
IHC 2/3 (*n* = 30)	13 (43.3) (95% CI: 25.5–62.6)	8 (26.7)	8 (26.7)
IHC 3 (*n* = 10)	5 (50.0) (95% CI: 22.2–77.8)	2 (20.0)	3 (30.0)
IHC 2 (*n* = 20)	8 (40.0) (95% CI: 20.9–63.9)	6 (30.0)	5 (25.0)
IHC 0/1 (*n* = 35)	4 (11.4) (95% CI: 4.0–26.3)	13 (37.1)	13 (37.1)
IHC 1 (*n* = 23)	3 (13.0) (95% CI: 3.7–31.7)	8 (34.8)	8 (34.8)
IHC 0 (*n* = 12)	1 (8.3) (95% CI: 0.4–34.9)	5 (41.7)	5 (41.7)

Figure 1 | PD-L1 prevalence and response rates in patients with UBC. **a**, PD-L1 prevalence by immunohistochemistry (IHC) in patients screened for the PCD4989g clinical trial. PD-L1 positivity was defined as ≥5% of tumour-infiltrating immune cells or tumour cells staining for PD-L1 by IHC. **b**, Representative images (×20 magnification) of PD-L1 IHC staining of tumours from patients with UBC. **c**, Response rates, including overall response, stable disease and progressive disease by tumour-infiltrating immune cell PD-L1 IHC status. Includes both confirmed and unconfirmed responses per RECIST v1.1. Five of seventeen responses were unconfirmed. Best response is not known for seven patients.

patients with IHC 2/3 tumours and a minimum of 12 weeks of follow-up, an ORR of 52% (13 of 25; 95% CI: 32–70%) was achieved. Sixteen of the seventeen responders had ongoing responses, and all seventeen responders continued on treatment with MPDL3280A at the data cutoff. One patient who initially responded at the first response assessment later presented with new lesions, including a bladder mass thought to be consistent with pseudoprogression. A biopsy of the new mass revealed extensive necrosis. This patient continued on treatment and had completed 12 cycles at the time of the data cutoff.

While the median has not been reached, duration of response ranged from 0.1+ to 30.3+ weeks for patients with IHC 2/3 tumours and from 0.1+ to 6.0+ weeks for patients with IHC 0/1 tumours. Furthermore, while response to MPDL3280A was associated with the tumour-infiltrating immune cell IHC scores ($P = 0.026$), there did not appear to be an association with tumour cell IHC scores ($P = 0.93$; Extended Data Table 3).

Exploratory subgroup analyses demonstrated that IHC 2/3 and IHC 0/1 patients with an ECOG performance score of 1 had ORRs of 33% (5 of 15) and 14% (3 of 22), respectively, while patients whose time from previous chemotherapy was ≤3 months had ORRs of 33% (3 of 9) and 19% (3 of 16), respectively. The ORRs for patients with visceral metastases at baseline was 21% (4 of 19) and 10% (3 of 29) for IHC 2/3 and IHC 0/1 patients, respectively, while the ORRs for patients with no visceral metastases at baseline were 82% (9 of 11) and 17% (1 of 6) for IHC 2/3 and IHC 0/1 patients, respectively. The ORRs in current/former smokers and never smokers were 25% (11 of 44) and 26% (6 of 23), respectively. In total, most patients (55%) had a reduction in tumour burden as measured by Response Evaluation Criteria in Solid Tumours, version 1.1 (RECIST v1.1) (Fig. 2b). Overall, responses were rapid and occurred at a median of 42 days from starting treatment (Fig. 2c). Patients with IHC 2/3 tumours and IHC 0/1 tumours had a median follow up of 4.2 months (range: 1.1+ to 8.5 months) and 2.7 months (range: 0.7+ to 3.6 months), respectively. Twenty-five patients (37%) had been discontinued from the study due to disease progression ($n = 17$), death ($n = 4$), lost to follow-up ($n = 1$), physician decision ($n = 2$), or patient decision ($n = 1$).

Over the course of treatment with MPDL3280A, cytokines and circulating cells were monitored. Transient elevations in cytokines, including interleukin (IL)-18 and interferon (IFN)-γ, were observed by cycle 2 day 1. A similar dynamic profile was observed for proliferating $CD8^+$ $HLA\text{-}DR^+Ki\text{-}67^+$ T cells (Extended Data Fig. 3), consistent with the MPDL3280A mechanism of action. These markers were altered in all patients treated and were not associated with response.

There is an urgent need for efficacious and well-tolerated therapies in metastatic UBC, as even first-line chemotherapy is poorly tolerated in a large proportion of individuals[14,15]. The study results presented here demonstrate that not only can MPDL3280A treatment achieve

Table 1 | Baseline characteristics of efficacy-evaluable patients with UBC

Characteristic	PD-L1 IHC 2/3 ($n = 30$)	PD-L1 IHC 0/1 ($n = 35$)	Efficacy-evaluable patients ($n = 67$)
Age (years)			
Median	66.5	63.0	65.0
Range	42–86	36–81	36–86
Sex (*n* (%))			
Male	25 (83.3)	21 (60.0)	48 (71.6)
ECOG PS (*n* (%))			
0	14 (48.3)*	13 (37.1)†	27 (40.9)‡
1	15 (51.7)*	22 (62.9)†	39 (59.1)‡
Smoking status (*n* (%))			
Current/previous smoker	20 (66.7)	22 (62.9)	44 (65.7)
Site of primary tumour (*n* (%))			
Bladder	27 (90.0)	28 (80.0)	55 (82.1)
Renal pelvis	1 (3.3)	3 (8.6)	4 (6.0)
Ureter	0	3 (8.6)	5 (7.5)
Urethra	2 (6.7)	1 (2.9)	3 (4.5)
Sites of metastases at baseline (*n* (%))			
Visceral	19 (63.3)	29 (82.9)	50 (74.6)
Liver	9 (30.0)	12 (34.3)	22 (32.8)
Prior treatments (*n* (%))			
Cystectomy	20 (66.7)	11 (31.4)	32 (47.8)
Chemotherapy	29 (96.7)	31 (88.6)	62 (92.5)
Prior platinum	29 (96.7)	31 (88.6)	62 (92.5)
Cisplatin	28 (93.3)	23 (65.7)	53 (79.1)
Carboplatin	7 (23.3)	15 (42.9)	23 (34.3)
≥2 Prior systemic regimens	21 (70.0)	25 (71.4)	48 (71.6)
Prior BCG	6 (20.0)	5 (14.3)	11 (16.4)
≤3 months from last prior chemotherapy (*n* (%))	9 (31.0)*	16 (51.6)§	26 (41.9)‖
Organ function (*n* (%))			
Alkaline phosphatase ≥ULN	4 (13.3)	10 (28.6)	16 (23.9)
CrCl <60 ml min^{-1}	7 (23.3)¶	13 (38.2)#	22 (33.3)‡
Haemoglobin <10 g dl^{-1}	2 (6.9)*	9 (26.5)#	12 (18.5)☆
PD-L1 IHC (*n* (%))			
0	0 (0)	12 (34.3)	12 (17.9)
1	0 (0)	23 (65.7)	23 (34.3)
2	20 (66.7)	0 (0)	20 (29.9)
3	10 (33.3)	0 (0)	10 (14.9)
Unknown	0 (0)	0 (0)	2 (3.0)

BCG, Bacille Calmette–Guerin; CrCl, creatinine clearance; ULN, upper limit of normal. Two patients have unknown IHC status.
* $n = 29$. † $n = 35$. ‡ $n = 66$. § $n = 31$; ‖ $n = 62$; ¶ $n = 30$; # $n = 34$; ☆ $n = 65$.

high response rates, but also that the likelihood of response can be increased by determining the PD-L1 status of tumour-infiltrating immune cells. Previous biomarker analysis with immune check point inhibitors has focused on PD-L1 expression on tumour cells rather than tumour-infiltrating immune cells. The observation that expression of immune infiltrates on pre-treatment tissue—which can be far removed temporally, anatomically and biologically from the metastatic tumours—correlated with outcomes provides some insight into the underlying stability of immune-related tumour surveillance in UBC. For example, the tissues examined here were originally obtained between 0 and 10 years before cycle 1, day 1, with most tissues being obtained within 4 years (Extended Data Fig. 1). The association of response to MPDL3280A with PD-L1 expression on tumour-infiltrating immune cells was also recently observed in lung cancer[9].

Table 2 | Treatment-related adverse events occurring in two or more patients (grade 1–2) or in one patient (grade 3–4)

Treatment-related adverse events* ($n = 68$)	All grades (*n* (%))	Grade 3–4 (*n* (%))
All	39 (57.4)	3 (4.4)
Decreased appetite	8 (11.8)	0
Fatigue	8 (11.8)	0
Nausea	8 (11.8)	0
Pyrexia	6 (8.8)	0
Asthenia	5 (7.4)	1 (1.5)
Chills	3 (4.4)	0
Influenza-like illness	3 (4.4)	0
Lethargy	3 (4.4)	0
Anaemia	2 (2.9)	0
Arthralgia	2 (2.9)	0
Bone pain	2 (2.9)	0
Hyperthermia	2 (2.9)	0
Pain	2 (2.9)	0
Platelet count decrease	2 (2.9)	0
Pruritus	2 (2.9)	0
Thrombocytopaenia	2 (2.9)	1 (1.5)
Vomiting	2 (2.9)	0
Blood phosphorus decrease	1 (1.5)	1 (1.5)

* National Cancer Institute Common Terminology Criteria for Adverse Events, version 4.0.

While cross-study comparisons are limited, the 43% (95% CI: 26–63%) response rate achieved here in patients with PD-L1 IHC 2/3 tumours provides evidence of noteworthy clinical activity of MPDL3280A in patients with UBC and compares favourably with that previously seen with single-agent salvage regimens[16–22]. In addition, patients with PD-L1 IHC 0/1 tumours had a response rate of 11% (95% CI: 4–26%), consistent with historic response rates of 9–11% in randomized studies for patients with relapsed metastatic UBC[1,2]. Responses in this heavily pre-treated population were also rapid and occurred in patients with poor prognostic features.

Chemotherapy is challenging to administer in patients with UBC who have a median age at diagnosis of 73 years and multiple co-morbidities[23] (see also http://seer.cancer.gov/statfacts/html/urinb.html). Many patients forgo chemotherapy for metastatic disease due to the toxicity and the limited durable benefit, and only approximately 40% of patients receive second-line treatment[23]. Therefore, the safety results with MPDL3280A are also encouraging. The larger and longer safety experience in the overall phase I study further indicates that MPDL3280A is well tolerated, with AE rates lower than many of the standard second-line treatment options for metastatic UBC[9].

To gain a better understanding of how the immune system responds to MPDL3280A, the levels of the IL-18 immunostimulatory cytokine and IFN-γ, which is stimulated by IL-18, were examined over several

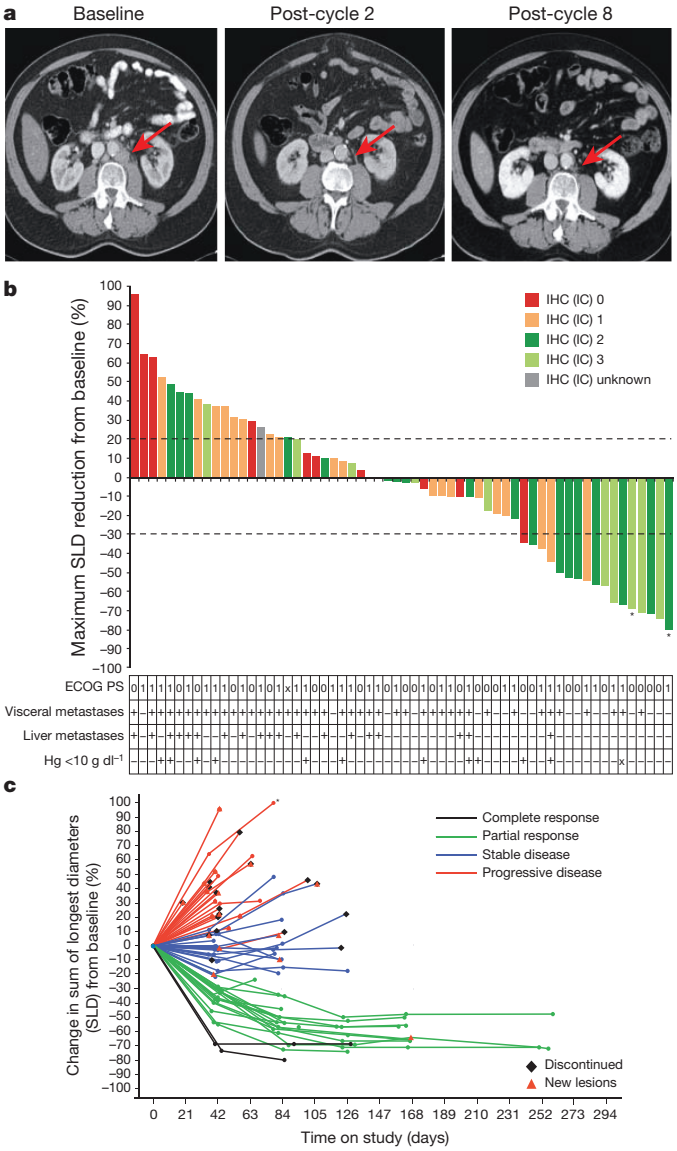


Figure 2 | MPDL3280A anti-tumour activity in patients with UBC. **a**, Example of a tumour response in a 68-year-old male who was initially treated for bladder cancer between 2011 and 2012. He underwent transurethral resection of the bladder tumour and intravesical Bacille Calmette–Guerin for localized recurrent disease (G3pT1c). Subsequently, cystectomy and adjuvant cisplatin-based chemotherapy were given for pT3aN2 disease. Nine months after the completion of chemotherapy, retroperitoneal lymph node relapse occurred. The patient refused further chemotherapy. As the patient was found to be PD-L1 positive, he was enrolled and started treatment with MPDL3280A on 18 July 2013. After two cycles, a computed tomography scan demonstrated complete remission. As of the data cutoff, he has received eight cycles of treatment without evidence of disease progression. (See Extended Data Fig. 2 for further details.) **b**, Waterfall plot according to tumour-infiltrating immune cell immunohistochemical (IHC) status, measuring the maximum percentage reduction from baseline in the sum of the longest diameter (SLD) for target lesions; +20% and −30% are marked by dashed lines. Eastern Cooperative Oncology Group (ECOG) performance status and the presence of visceral metastases, liver metastases, or haemoglobin (Hg) $<10\,g\,dl^{-1}$ are given by patient. X indicates missing data. Asterisk indicates patients with complete responses who had ≤100% reduction of the target lesions due to lymph node target lesions; all lymph nodes returned to normal size per Response Evaluation Criteria in Solid Tumours v1.1. **c**, Spaghetti plot providing response, presence of new lesions (red triangle) and discontinuation (black diamond) by patients over time. Figure does not include efficacy-evaluable patients with UBC who did not have any post-baseline tumour assessments. Asterisk indicates a value >100%.

cycles. IL-18 and IFN-γ levels transiently increased, in line with both having an important role in innate and adaptive immune responses[24] as well as functioning in the proliferation of naive and memory $CD8^+$ T cells[25]. Accordingly, a similar pattern was identified for $CD8^+HLA\text{-}DR^+Ki\text{-}67^+$ cells. As these changes occurred in all patients receiving MPDL3280A, they are indicative of a potential systemic host response to PD-L1 pathway inhibition and could provide a non-invasive immune monitoring tool. These dynamic, but transient, changes in the blood do not necessarily reflect the expression of immune parameters within the tumours, and examining sequential tissue during treatment with MPDL3280A will provide more insight into the molecular responses of tumours to MPDL3280A.

Cancers with a high rate of somatic mutations, including non-small cell lung cancer, melanoma and UBC, appear to respond well to MPDL3280A. One hypothesis that explains this result is that patients with these cancers have an increase in tumour-specific antigens[4,5]. Further work to evaluate the frequency of somatic mutations at baseline will help to elucidate the relationship between mutational frequency and response to PD-L1 blockade.

This study provides striking preliminary efficacy and safety results with MPDL3280A for the treatment of UBC. Additionally, our data demonstrate the potential of immune cell PD-L1 levels as a biomarker. Our trial employed an adaptive-type design instead of a traditional phase I approach with a fixed sample size. Using this approach for the expansion stage (details are provided in Methods) rapidly identifies and characterizes monotherapy activity in tumours types for which there are no expected spontaneous responses, including UBC. A futility-type rule was applied within each indication to suspend enrolment in that indication if there were no responders observed by a certain enrolment number. Additionally, expansion cohorts could be enrolled to achieve certain precision in the safety and response rate estimates. This approach has been particularly useful for therapies that have rapid and strong monotherapy activity in a broad range of cancer types.

Many recent phase I trials have used a similar adaptive-type design approach, without explicit power and type I error considerations[26,27]. This design allows for the exploration of the frequency and relevance of biomarkers as well as the rapid assessment of efficacy in specific tumour types. These trials, including ours, tend to recruit relatively large numbers of patients with specific characteristics (tumour types and biomarkers) into the expansion cohorts to increase the precision of the results. The flexibility that results from this type of trial design can be helpful when planning prospective randomized trials.

On the basis of these data, the FDA granted MPDL3280A breakthrough status for UBC. Further investigation of MPDL3280A in UBC is warranted, in multiple settings, including in patients who have failed or are intolerable towards initial chemotherapy. Clinical studies are enrolling patients to study MPDL3280A in bladder and other cancers.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 23 May; accepted 30 September 2014.

1. Choueiri, T. K. *et al.* Double-blind, randomized trial of docetaxel plus vandetanib versus docetaxel plus placebo in platinum-pretreated metastatic urothelial cancer. *J. Clin. Oncol.* **30,** 507–512 (2012).
2. Bellmunt, J. *et al.* Phase III trial of vinflunine plus best supportive care compared with best supportive care alone after a platinum-containing regimen in patients with advanced transitional cell carcinoma of the urothelial tract. *J. Clin. Oncol.* **27,** 4454–4461 (2009).
3. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507,** 315–322 (2014).
4. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).
5. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502,** 333–339 (2013).
6. Chen, D. S. & Mellman, I. Oncology meets immunology: the cancer-immunity cycle. *Immunity* **39,** 1–10 (2013).
7. Chen, D. S., Irving, B. A. & Hodi, F. S. Molecular pathways: next-generation immunotherapy—inhibiting programmed death-ligand 1 and programmed death-1. *Clin. Cancer Res.* **18,** 6580–6587 (2012).

8. van Rooij, N. *et al.* Tumor exome analysis reveals neoantigen-specific T-cell reactivity in an ipilimumab-responsive melanoma. *J. Clin. Oncol.* **31,** e439–e442 (2013).
9. Herbst, R. S. *et al.* Predictive correlates of response to anti-PD-L1 in cancer patients. *Nature* http://dx.doi.org/10.1038/nature14011 (this issue).
10. Topalian, S. L. *et al.* Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *N. Engl. J. Med.* **366,** 2443–2454 (2012).
11. Bellmunt, J. *et al.* Prognostic factors in patients with advanced transitional cell carcinoma of the urothelial tract experiencing treatment failure with platinum-containing regimens. *J. Clin. Oncol.* **28,** 1850–1855 (2010).
12. Sonpavde, G. *et al.* Time from prior chemotherapy enhances prognostic risk grouping in the second-line setting of advanced urothelial carcinoma: a retrospective analysis of pooled, prospective phase 2 trials. *Eur. Urol.* **63,** 717–723 (2013).
13. Wrammert, J. *et al.* Rapid cloning of high-affinity human monoclonal antibodies against influenza virus. *Nature* **453,** 667–671 (2008).
14. De Santis, M. *et al.* Randomized phase II/III trial assessing gemcitabine/carboplatin and methotrexate/carboplatin/vinblastine in patients with advanced urothelial cancer who are unfit for cisplatin-based chemotherapy: EORTC study 30986. *J. Clin. Oncol.* **30,** 191–199 (2012).
15. Dreicer, R., Gustin, D. M., See, W. A. & Williams, R. D. Paclitaxel in advanced urothelial carcinoma: its role in patients with renal insufficiency and as salvage therapy. *J. Urol.* **156,** 1606–1608 (1996).
16. Gallagher, D. J. *et al.* Phase II study of sunitinib in patients with metastatic urothelial cancer. *J. Clin. Oncol.* **28,** 1373–1379 (2010).
17. Necchi, A. *et al.* Pazopanib in advanced and platinum-resistant urothelial cancer: an open-label, single group, phase 2 trial. *Lancet Oncol.* **13,** 810–816 (2012).
18. Seront, E. *et al.* Phase II study of everolimus in patients with locally advanced or metastatic transitional cell carcinoma of the urothelial tract: clinical activity, molecular response, and biomarkers. *Ann. Oncol.* **23,** 2663–2670 (2012).
19. Vaughn, D. J. *et al.* Vinflunine in platinum-pretreated patients with locally advanced or metastatic urothelial carcinoma: results of a large phase 2 study. *Cancer* **115,** 4110–4117 (2009).
20. Sweeney, C. J. *et al.* Phase II study of pemetrexed for second-line treatment of transitional cell cancer of the urothelium. *J. Clin. Oncol.* **24,** 3451–3457 (2006).
21. Ko, Y. J. *et al.* Nanoparticle albumin-bound paclitaxel for second-line treatment of metastatic urothelial carcinoma: a single group, multicentre, phase 2 study. *Lancet Oncol.* **14,** 769–776 (2013).
22. Culine, S. *et al.* A phase II study of vinflunine in bladder cancer patients progressing after first-line platinum-containing regimen. *Br. J. Cancer* **94,** 1395–1401 (2006).
23. Chen, G. J., Galsky, M. D., Latini, D. M., Sonpavde, G. & DeBakey, M. E. Patterns of chemotherapy and survival in elderly patients with advanced bladder cancer: a large Medicare database study. *J. Clin. Oncol.* (suppl.) abstr. 4551 (2013).
24. Okamura, H. *et al.* Cloning of a new cytokine that induces IFN-γ production by T cells. *Nature* **378,** 88–91 (1995).
25. Iwai, Y. *et al.* An IFN-γ-IL-18 signaling loop accelerates memory $CD8^+$ T cell proliferation. *PLoS ONE* **3,** e2404 (2008).
26. Topalian, S. L. *et al.* Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *N. Engl. J. Med.* **366,** 2443–2454 (2012).
27. Dahlberg, S. E., Shapiro, G. I., Clark, J. W. & Johnson, B. E. Evaluation of statistical designs in phase I expansion cohorts: the Dana-Farber/Harvard Cancer Center experience. *J. Natl Cancer Inst.* **106,** dju163 (2014).

Acknowledgements We thank the patients and their families. Additionally, we thank the investigators and their staff, including the Barts Health NHS Trust and the Royal Free Foundation Trust, A. Balmanoukian and O. Hamid (The Angeles Clinic and Research Institute), J. Powderly (Carolina BioOncology Institute), P. Cassier (Centre Léon-Bérard), F. Steven Hodi (Dana-Farber Cancer Institute), J.-C. Soria (Gustave Roussy), J. P. DeLord (Institute Claudius Regaud), C. Drake and L. Emens (Johns Hopkins), D. Lawrence and R. Lee (Massachusetts General Hospital), S. Antonia and J. Zhang (Moffitt Cancer Center), M. Gordon (Pinnacle Oncology Hematology), H. Kohrt and S. Srinivas (Stanford University Cancer Institute), and J. Tabernero (Vall d'Hebron University Hospital). Support for third-party writing assistance for this manuscript was provided by F. Hoffmann-La Roche Ltd.

Author Contributions T.P., G.D.F., D.P.P., D.S.C. and N.J.V. contributed to the overall study design; Z.B. and P.S.H. provided the biomarker studies; S.-I.T. performed the statistical analysis. All authors analysed the data. All authors contributed to writing the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.P. (Thomas.Powles@bartshealth.nhs.uk).

LETTER

doi:10.1038/nature14011

Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients

Roy S. Herbst[1], Jean-Charles Soria[2], Marcin Kowanetz[3], Gregg D. Fine[3], Omid Hamid[4], Michael S. Gordon[5], Jeffery A. Sosman[6], David F. McDermott[7], John D. Powderly[8], Scott N. Gettinger[1], Holbrook E. K. Kohrt[9], Leora Horn[10], Donald P. Lawrence[11], Sandra Rost[3], Maya Leabman[3], Yuanyuan Xiao[3], Ahmad Mokatrin[3], Hartmut Koeppen[3], Priti S. Hegde[3], Ira Mellman[3], Daniel S. Chen[3] & F. Stephen Hodi[12]

The development of human cancer is a multistep process characterized by the accumulation of genetic and epigenetic alterations that drive or reflect tumour progression. These changes distinguish cancer cells from their normal counterparts, allowing tumours to be recognized as foreign by the immune system[1–4]. However, tumours are rarely rejected spontaneously, reflecting their ability to maintain an immunosuppressive microenvironment[5]. Programmed death-ligand 1 (PD-L1; also called B7-H1 or CD274), which is expressed on many cancer and immune cells, plays an important part in blocking the 'cancer immunity cycle' by binding programmed death-1 (PD-1) and B7.1 (CD80), both of which are negative regulators of T-lymphocyte activation. Binding of PD-L1 to its receptors suppresses T-cell migration, proliferation and secretion of cytotoxic mediators, and restricts tumour cell killing[6–10]. The PD-L1–PD-1 axis protects the host from overactive T-effector cells not only in cancer but also during microbial infections[11]. Blocking PD-L1 should therefore enhance anticancer immunity, but little is known about predictive factors of efficacy. This study was designed to evaluate the safety, activity and biomarkers of PD-L1 inhibition using the engineered humanized antibody MPDL3280A. Here we show that across multiple cancer types, responses (as evaluated by Response Evaluation Criteria in Solid Tumours, version 1.1) were observed in patients with tumours expressing high levels of PD-L1, especially when PD-L1 was expressed by tumour-infiltrating immune cells. Furthermore, responses were associated with T-helper type 1 (T_H1) gene expression, CTLA4 expression and the absence of fractalkine (CX3CL1) in baseline tumour specimens. Together, these data suggest that MPDL3280A is most effective in patients in which pre-existing immunity is suppressed by PD-L1, and is re-invigorated on antibody treatment.

Pre-clinical studies demonstrated that anti-PD-L1 treatment of mice bearing implanted syngeneic tumours could lead to tumour regression and the induction of protective immune memory in the setting of re-challenge with tumour cells (Genentech, unpublished data). However, most mouse models constitutively express PD-L1 (ref. 12), which is not consistent with human tumours. Additionally, only a few syngeneic models (notably the MC38 colon carcinoma model) were responsive to anti-PD-L1 as a single agent (Genentech, unpublished data). Therefore, a detailed analysis of PD-L1 expression in human tumours and its association with clinical benefit was required.

PD-L1 in human cancers was investigated using an anti-PD-L1 immunohistochemistry (IHC) antibody optimized for staining of formalin-fixed paraffin-embedded tissue samples. Staining of pre-treatment specimens submitted for our clinical study demonstrated expression across a range of cancers (Fig. 1a). PD-L1 staining was observed on tumour cells, as well as on tumour-infiltrating immune cells (Fig. 1b), with PD-L1-positive tumour-infiltrating immune cells being more common than PD-L1-positive tumour cells. PD-L1-positive tumour-infiltrating immune cells included myeloid cells (macrophages, dendritic cells) and T cells; B cells were negative for PD-L1 (Fig. 1c).

We developed a high-affinity human monoclonal immunoglobulin-G1 (IgG1) antibody for clinical use that specifically binds to PD-L1 (MPDL-3280A; binding affinity K_d (dissociation constant) = 0.4 nM) and prevents its interaction with PD-1 and B7.1. However, the antibody would leave intact the interaction of PD-1 with its alternative ligand PD-L2 (also called B7-DC or CD273), which is thought to have a key role in maintaining peripheral tolerance, particularly in the lung[13,14]. MPDL3280A was engineered with a crystallizable fragment (Fc) domain modification eliminating antibody-dependent cellular cytotoxicity at clinically relevant doses, preventing depletion of activated T cells[15,16] (see Methods).

Patients were treated with MPDL3280A, and pre-treatment and on-treatment tumour specimens were characterized from available samples. A total of 277 patients with advanced incurable cancer received MPDL-3280A intravenously every 3 weeks (q3w; Extended Data Fig. 1a, b and Extended Data Table 1; see Methods). Mean single-dose MPDL3280A pharmacokinetics were consistent with a typical IgG1 at doses $\geq$1 mg kg^{-1}, with a mean terminal serum half-life of $\sim$3 weeks (Extended Data Fig. 1c). Overall, treatment was well tolerated up to the maximum administered dose of 20 mg kg^{-1} q3w (Table 1).

Most adverse events (AEs) did not require medical treatment. The most common treatment-related AE was fatigue (Table 1), which often occurred with low-grade fever during the first treatment cycle. Pyrexia was reported in $\sim$21% of patients; it most commonly occurred during cycle 1 and was uncommon during subsequent cycles (Extended Data Fig. 2a). Additionally, an $\sim$2-fold increase in activated proliferating CD8$^+$ T cells (CD8$^+$HLA-DR$^+$Ki-67$^+$) and a trend of increased circulating interferon (IFN)-γ were observed by the end of the first cycle (Extended Data Fig. 2b, c).

Treatment-related grade 3–4 AEs were observed in 35 patients (13%) and immune-related grade 3–4 AEs were observed in 3 patients (1%) (see Methods for further information regarding AE grades). No cases of grades 3–5 pneumonitis were seen.

The impact of PD-L1 inhibition on metastatic lesions was evaluated per Response Evaluation Criteria in Solid Tumours, version 1.1 (RECIST v1.1). In the 175 efficacy-evaluable patients (with demographic and baseline characteristics similar to those in all patients), confirmed responses (complete and partial responses) were observed in 32 of 175 (18%), 11 of 53 (21%), 11 of 43 (26%), 7 of 56 (13%) and 3 of 23 (13%) of patients with all tumour types, non-small cell lung cancer (NSCLC), melanoma, renal cell carcinoma and other tumours (including colorectal cancer, gastric cancer, and head and neck squamous cell carcinoma), respectively.

[1]Yale Comprehensive Cancer Center, Yale School of Medicine, 333 Cedar Street, WWW221, New Haven, Connecticut 06520, USA. [2]Gustave Roussy South-Paris University, 114 Rue Edouard Vaillant, 94805 Villefuij, Cedex, France. [3]Genentech, Inc., 1 DNA Way, South San Francisco, California 94080, USA. [4]The Angeles Clinic and Research Institute, 11818 Wilshire Blvd, Los Angeles, California 90025, USA. [5]Pinnacle Oncology Hematology, 9055 E Del Camino Dr 100, Scottsdale, Arizona 85258, USA. [6]Vanderbilt-Ingram Cancer Center, 2220 Pierce Avenue, Nashville, Tennessee 37212, USA. [7]Beth Israel Deaconess Medical Center, 330 Brookline Avenue, Shapiro 9, Boston, Massachusetts 02215, USA. [8]Carolina BioOncology Institute, 9801 W. Kincey Ave, Suite 145, Huntersville, North Carolina 28078, USA. [9]Stanford University, CCSR Bldg Room 1110, Stanford, California 94305, USA. [10]Vanderbilt-Ingram Cancer Center, 1301 Medical Center Dr, Suite 1710, Nashville, Tennessee 37212, USA. [11]Massachusetts General Hospital, 55 Fruit Street, YAW 9E, Boston, Massachusetts 02114, USA. [12]Dana-Farber/Brigham and Women's Cancer Center, 450 Brookline Avenue, Boston, Massachusetts 02215, USA.

a PD-L1 prevalence determined with a Genentech/Roche anti-PD-L1 IHC assay

Indication	*n*	Percentage of PD-L1 positive (IC)	Percentage of PD-L1 positive (TC)
NSCLC	184	26	24
RCC	88	25	10
Melanoma	58	36	5
HNSCC	101	28	19
Gastric cancer	141	18	5
CRC	77	35	1
Pancreatic cancer	83	12	4

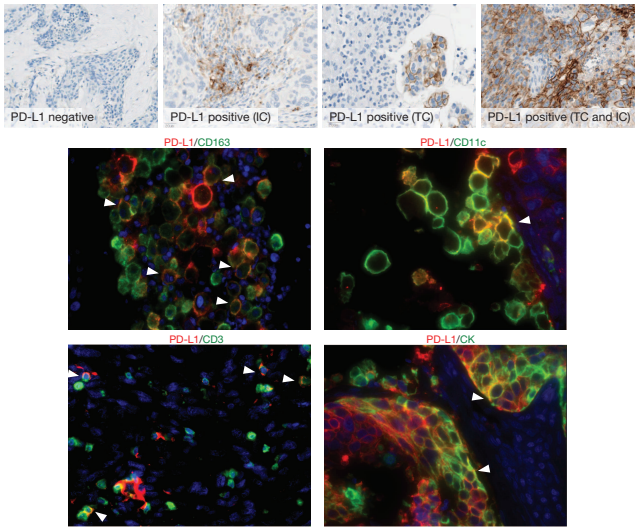


Figure 1 | Programmed death-ligand 1 (PD-L1) prevalence and expression. **a**, PD-L1 prevalence by immunohistochemistry (IHC) in samples collected for PCD4989g. PD-L1 positivity was defined as ≥5% of tumour-infiltrating immune cells (ICs) or tumour cells (TCs) staining for PD-L1 by IHC. **b**, Representative images of PD-L1 by IHC (brown) in tumours from patients with non-small cell lung cancer (NSCLC). The PD-L1-negative image is at 20× magnification, other images at 40× magnification. **c**, Co-localization of PD-L1 with selected tumour-infiltrating immune cell and tumour cell markers by immunofluorescence in NSCLC and melanoma tumours. PD-L1 staining in red; markers of tumour-infiltrating immune cells and tumour cells in green; and DAPI staining in blue. Areas of overlap are indicated with white arrowheads. All four images are at 40× resolution. Markers of tumour-infiltrating immune cells: CD163 (macrophages), CD11c (dendritic cells) and CD3 (T cells). Marker of tumour cells: cytokeratin (CK). CRC, colorectal cancer; HNSCC, head and neck squamous cell carcinoma; NSCLC, non-small cell lung cancer; RCC, renal cell carcinoma.

Four more patients had unconfirmed responses (Table 2, Fig. 2 and Extended Data Fig. 3a). Responses could also be rapid and durable (Fig. 2b and Extended Data Fig. 3a), with shrinking or resolving palpable lesions detected within days in some responders and nearly all responders (especially patients with NSCLC) continuing to respond and staying on study. In addition, RECIST may not accurately describe the full spectrum of responses observed because some patients who had a best response of progressive disease per RECIST went on to develop durable tumour shrinkage or prolonged stable disease (pseudoprogression)[17]. The median progression-free survival of all patients was 18 weeks. We also performed an exploratory analysis of patients with NSCLC and detected a potential trend of former/current smokers responding better to MPDL3280A versus never smokers (11 of 26 (42%) versus 1 of 10 (10%), respectively; $P = 0.4229$ using a Fisher exact test; see the accompanying paper (ref. 18) for further discussion).

There appears to be an association between response and the expression of PD-L1 in pre-treatment samples (Fig. 3 and Extended Data Figs 3 and 4). The association of response to MPDL3280A treatment and tumour-infiltrating immune cell PD-L1 expression reached statistical significance (NSCLC, $P = 0.015$ (Fig. 3a and Extended Data Fig. 4a); all tumours, $P = 0.007$ (Fig. 3b, c and Extended Data Fig. 4b)), while the association with tumour cell PD-L1 expression did not (NSCLC, $P = 0.920$ (Extended Data Fig. 4c); all tumours, $P = 0.079$ (Extended Data Fig. 4d)).

Table 1 | Adverse events

Treatment-related AEs (*n* = 277)			AEs regardless of attribution (*n* = 277)		
Events (≥4% of patients)	Any grade (*n* (%))	Grade 3–4 (*n* (%))	Events (≥5% of patients)	Any grade (*n* (%))	Grade 3–4 (*n* (%))
Any AE	194 (70.0)	35 (12.6)	Any AE	263 (94.9)	108 (39.0)
Fatigue	67 (24.2)	5 (1.8)	Fatigue	100 (36.1)	9 (3.2)
Decreased appetite	33 (11.9)	–	Nausea	69 (24.9)	2 (0.7)
Nausea	32 (11.6)	1 (0.4)	Dyspnoea	66 (23.8)	11 (4.0)
Pyrexia	32 (11.6)	–	Decreased appetite	64 (23.1)	–
Diarrhoea	29 (10.5)	–	Cough	60 (21.7)	–
Rash	29 (10.5)	–	Diarrhoea	60 (21.7)	–
Pruritus	23 (8.3)	–	Pyrexia	57 (20.6)	–
Arthralgia	22 (7.9)	–	Constipation	55 (19.9)	–
Headache	21 (7.6)	1 (0.4)	Headache	49 (17.7)	–
Chills	19 (6.9)	–	Vomiting	46 (16.6)	–
Influenza-like illness	16 (5.8)	1 (0.4)	Anaemia	44 (15.9)	10 (3.6)
Asthenia	15 (5.4)	2 (0.7)	Insomnia	43 (15.5)	–
Dyspnea	15 (5.4)	2 (0.7)	Back pain	42 (15.2)	4 (1.4)
Pain	15 (5.4)	1 (0.4)	Arthralgia	41 (14.8)	–
Myalgia	13 (4.7)	–	Rash	40 (14.4)	–
Anaemia	12 (4.3)	2 (0.7)	Asthenia	34 (12.3)	4 (1.4)
Dry skin	12 (4.3)	–	Pruritus	33 (11.9)	–
Night sweats	12 (4.3)	–	Chills	31 (11.2)	–
Vomiting	11 (4.0)	1 (0.4)	Upper respiratory tract infection	30 (10.8)	–
Other grade 3–4 AEs, ≥2 patients			Anxiety	20 (7.2)	–
ALT increased	6 (2.2)	3 (1.1)	Influenza-like illness	20 (7.2)	–
AST increased	4 (1.4)	3 (1.1)	Nasal congestion	20 (7.2)	–
Hypoxia	4 (1.4)	3 (1.1)	Urinary tract infection	20 (7.2)	3 (1.1)
Hyperglycaemia	4 (1.4)	2 (0.7)	Dehydration	19 (6.9)	4 (1.4)
Hyponatraemia	4 (1.4)	2 (0.7)	Hyperglycaemia	19 (6.9)	7 (2.5)
Cardiac tamponade	2 (0.7)	2 (0.7)	Myalgia	19 (6.9)	–
Hypophosphataemia	2 (0.7)	2 (0.7)	Night sweats	19 (6.9)	–
Tumour lysis syndrome	2 (0.7)	2 (0.7)	Productive cough	19 (6.9)	–
			Dry skin	16 (5.8)	–
			Dry mouth	14 (5.1)	–
			Hypoxia	14 (5.1)	6 (2.2)
			Weight decreased	14 (5.1)	–

AE, adverse event; ALT, alanine aminotransferase; AST, aspartate aminotransferase.

Table 2 | Efficacy of MPDL3280A across tumour types

Tumour types	ORR* (*n* (%)) (95% CI)	SD (best response) (*n* (%))	PD (best response) (*n* (%))	SD ≥24 weeks (*n* (%))	24-week PFS (%)
Overall (*n* = 175)	36 (21) (15–27)	68 (39)	65 (37)	33 (19)	42
NSCLC (*n* = 53)	12 (23) (12–35)	18 (34)	21 (40)	9 (17)	45
Non-squamous (*n* = 42)	9 (21) (11–36)	16 (38)	16 (38)	7 (17)	44
Squamous (*n* = 11)	3 (27) (8–61)	2 (18)	5 (46)	2 (18)	46
Melanoma (*n* = 43)	13 (30) (18–45)	11 (26)	18 (42)	4 (9)	41
Cutaneous (*n* = 33)	12 (36) (21–55)	9 (27)	11 (33)	4 (12)	51
Mucosal (*n* = 5)	1 (20) (1–66)	0	4 (80)	0	20
Ocular (*n* = 4)	0 (0–53)	1 (25)	3 (75)	0	0
RCC (*n* = 56)	8 (14) (6–25)	30 (54)	17 (30)	18 (32)	48
Clear cell (*n* = 49)	7 (14) (7–27)	28 (57)	14 (29)	18 (37)	52
Non-clear cell (*n* = 7)	1 (14) (1–55)	2 (29)	3 (43)	0	17
Other (for example, CRC, GC and HNSCC; *n* = 23)†	3 (13) (4–32)	9 (39)	9 (39)	2 (9)	24

Patients dosed by 1 October 1 2012, with ≥1 mg kg^{-1} with a baseline tumour assessment. Data cutoff was 30 April 2013. CI, confidence interval; CRC, colorectal cancer; GC, gastric cancer; HNSCC, head and neck squamous cell carcinoma; NSCLC, non-small cell lung cancer; ORR, objective response rate; PD, progressive disease; PFS, progression-free survival; RCC, renal cell carcinoma; SD, stable disease.
* Per RECIST v1.1. All responses were confirmed except for in one patient with NSCLC, one patient with RCC and two patients with melanoma.
† Sarcoma (*n* = 2), ovarian (*n* = 1), head and neck (*n* = 6), breast (*n* = 3), colorectal (*n* = 6), pancreatic (*n* = 1), gastric (*n* = 1), oesophageal (*n* = 1), uterine (*n* = 1) and pancreatoduodenal (*n* = 1).

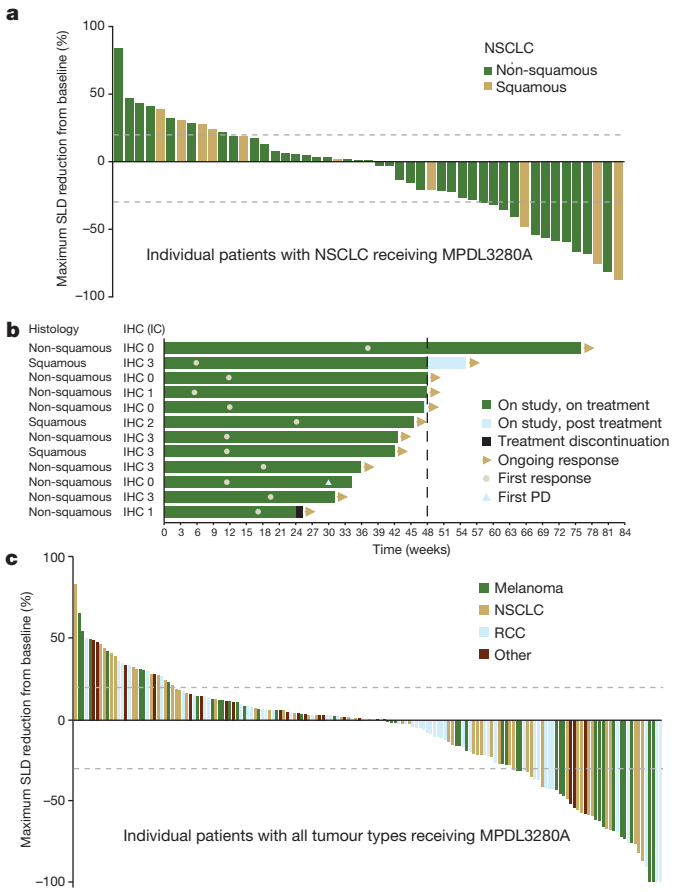


Figure 2 | Antitumour activity of MPDL3280A. **a**, A waterfall plot of patients with non-small cell lung cancer (NSCLC) measuring the maximum reduction from baseline in the sum of the longest diameter (SLD) for target lesions; +20% and −30% are marked by dashed lines. **b**, The time to response (Response Evaluation Criteria in Solid Tumours version 1.1) and the duration of study treatment for patients with NSCLC. The patient with progressive disease (PD) experienced ongoing clinical benefit as judged by the investigator. All but one response was confirmed. **c**, A waterfall plot of patients with all tumour types measuring the maximum reduction from baseline in the SLD for target lesions; +20% and −30% are marked by dashed lines. IC, tumour-infiltrating immune cells; IHC, immunohistochemistry; RCC, renal cell carcinoma.

For example, 83% of patients with IHC score 3 (tumour-infiltrating immune cell) NSCLC responded to treatments with only 17% progressing, whereas 43% of patients with IHC 2 (tumour-infiltrating immune cell) NSCLC were limited to disease stabilization (Fig. 3a and Extended Data Fig. 4a; see Methods for the IHC score definitions). Of the patients with IHC 3 (tumour cell) NSCLC, only 38% (3 of 8) responded while 38% (3 of 8) progressed (Extended Data Fig. 4c). There was also a trend between tumour IHC status and median progression-free survival (Fig. 3c).

When tumour samples were examined for the expression of different immune inhibitory factors (see Methods), the expected correlation with lack of response to MPDL3280A was not seen (Extended Data Fig. 5a, left panel). Instead there was a trend towards increased response in PD-L1-positive patients expressing a second negative regulator (Extended Data Fig. 5a, right panel). High PD-L2 expression did not appear to be associated with resistance to MPDL3280A, and some patients whose pre-treatment tumour biopsies showed the highest levels of PD-L2 expression experienced strong responses to MPDL3280A (for example, maximum sum of the longest diameter (SLD) decreases of 57%, 41% and 49%). Finally, the expression of CTLA4 and fractalkine in pre-treatment tumours appeared to correlate strongly with either response (CTLA4) or progression (fractalkine) after MPDL3280A (Extended Data Fig. 5b).

We compared results obtained for pre-treatment NSCLC tumours with those for renal cell carcinoma and melanoma (Extended Data Fig. 6). Although the expression of PD-L1 in MPDL3280A-responsive patients was a common feature, other aspects of the immune microenvironment appeared different. In melanoma, pre-treatment tumours in responding patients demonstrated elevated expression of IFN-γ as well as IFN-γ-inducible genes (for example, *IDO1* and *CXCL9*). These associations were weaker in patients with NSCLC or renal cell carcinoma.

To characterize the immunological events associated with tumour response or progression, serial on-treatment tumour biopsies were performed in 28 patients (Fig. 4a). After treatment, regressing lesions displayed a dense immune infiltrate and extensive tumour cell necrosis accompanied by the apparent sterilization of cancer cells in some cases (Extended Data Fig. 7a, b). A decrease in tumour SLD appeared to be accompanied by an increase in PD-L1 expression on tumour-infiltrating immune cells and tumour cells (Fig. 4a). The increase in PD-L1 expression with treatment correlated with changes in tumour IFN-γ expression (Pearson correlation coefficient = 0.70; Extended Data Fig. 5c). In addition, RNA isolated from regressing lesions was analysed for the presence of transcripts of immunological importance using a Fluidigm-based 'immunochip' (iChip, see Methods), and displayed expression patterns indicative of a generalized activation of CD8 and T_H1 T-cell responses (Extended Data Fig. 7c).

In contrast, most progressing patients with on-treatment biopsies showed a lack of PD-L1 upregulation by either tumour cells or tumour-infiltrating immune cells. These growing tumours displayed one of three patterns: (1) little or no tumour-infiltrating immune cell infiltration ('immunological ignorance'; Fig. 4b and Extended Data Fig. 8a); (2) presence of an intra-tumoral immune infiltrate with minimal to no expression of PD-L1 ('non-functional immune response'; Fig. 4b and Extended Data Fig. 8b); or (3) presence of an immune infiltrate that resided solely around the outer edge of the tumour cell mass ('excluded infiltrate'; Fig. 4b and Extended Data Fig. 9). Chip analysis of samples from these non-responders failed to provide evidence of activated T cells (Extended Data Figs 8b and 9). In cases where an excluded infiltrate of $CD8^+$ T cells was observed before treatment, PD-L1 inhibition did not induce

a Activity of MPDL3280A in NSCLC IHC (IC)

Diagnostic population	ORR (RECIST) (*n* (%))	SD (best response) (*n* (%))	SD ≥24 weeks (*n* (%))	PD (best response) (*n* (%))	24-week PFS (%)	Median PFS (weeks) (95% CI)
IHC 3 (*n* = 6)	5 (83)	0	0	1 (17)	83.3	NE (5,NE)
IHC 2 (*n* = 7)	1 (14)	3 (43)	0	2 (29)	14.3	11 (1,17)
IHC 1 (*n* = 13)	2 (15)	3 (23)	1 (8)	7 (54)	25.6	6 (5,43)
IHC 0 (*n* = 20)	4 (20)	7 (35)	4 (20)	9 (45)	45.0	13 (6,37)
Unknowns (*n* = 7)	0	5 (71)	4 (57)	2 (29)	71.4	NE (6,NE)
All patients (*n* = 53)	12 (23)	18 (34)	9 (17)	21 (40)	44.7	15 (6,43)

b Activity of MPDL3280A in all tumour types IHC (IC)

Diagnostic population	ORR (RECIST) (*n* (%))	SD (best response) (*n* (%))	SD ≥24 weeks (*n* (%))	PD (best response) (*n* (%))	24-week PFS (%)	Median PFS (weeks) (95% CI)
IHC 3 (*n* = 33)	15 (46)	9 (27)	4 (12)	8 (24)	60.0	37 (18,59)
IHC 2 (*n* = 23)	4 (17)	12 (52)	6 (26)	6 (26)	43.0	18 (6,48)
IHC 1 (*n* = 34)	7 (21)	11 (32)	6 (18)	14 (41)	40.9	17 (6,43)
IHC 0 (*n* = 60)	8 (13)	22 (37)	11 (18)	29 (48)	33.9	8 (6,18)
Unknowns (*n* = 25)	2 (8)	14 (56)	6 (24)	8 (32)	38.9	20 (6,NE)
All patients (*n* = 175)	36 (21)	68 (39)	33 (19)	65 (37)	42.2	18 (12,24)

c

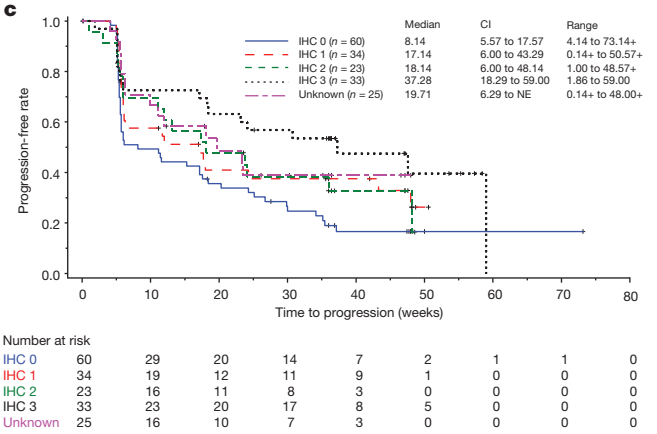


Figure 3 | Antitumour activity of MPDL3280A by immunohistochemistry (IHC) tumour-infiltrating immune cell (IC) and biomarker status. a, Table of antitumour activity in patients with NSCLC by PD-L1 IHC (IC) status. Patients with no post-first dose assessment were not estimable (NE; 1 with IHC 1 and 1 with IHC 2), but were included in the denominator for calculating objective response rate (ORR). **b**, Table of antitumour activity in patients with all tumour types by PD-L1 IHC (IC) status. Patients with no post-first dose assessment were not estimable and not included in the table (1 with IHC 0, 2 with IHC 1, 1 with IHC 2 and 1 with IHC 3), but were included in the denominator for calculating ORR. **c**, Kaplan–Meier curve showing the phase I percentage of progression-free survival by patient IHC (IC) status. Censored data are indicated by vertical tick marks. CI, confidence interval; NSCLC, non-small cell lung cancer; ORR, objective response rate; PD, progressive disease; PFS, progression free survival; PR, partial response; RECIST, Response Evaluation Criteria in Solid Tumours; SD, stable disease.

infiltration, although both proliferation and PD-L1 expression were detected in tumour-infiltrating immune cells at the tumour margin (Fig. 4b). Non-functional immune responses may explain why the presence of pre-treatment $CD8^+$ T cells in tumours (as opposed to the presence of PD-L1-positive infiltrates) failed to predict responses to MPDL3280A (Fig. 4b and Extended Data Fig. 8b).

Tumours that were non-responsive to MPDL3280A also did not exhibit an upregulation of genes associated with enhanced T-effector-cell activity in contrast to MPDL3280A-responsive tumours. Additionally, the expression of FOXP3 neither increased nor decreased in responding lesions, suggesting that T-regulatory cells may not have a major role in anti-PD-L1-responsive tumours.

Blood-based immune biomarkers were also examined. Several changes were observed, but these did not track significantly with response or progression following MPDL3280A administration. Increases in IL-18, ITAC (also called CXCL11 or IP-9) and $CD8^+HLA\text{-}DR^+Ki\text{-}67^+$ T cells, as well as a modest increase in IFN-γ, were observed during the first cycle of treatment (Extended Data Fig. 2b, c), whereas the average IL-6 expression levels exhibited a downward trend by cycle 2, day 1.

a Summary of responses to MPDL3280A in paired biopsies

	Increase in PD-L1 (TC) (no. (%))	Increase in PD-L1 (IC) (no. (%))
Maximum SLD decrease		
>30% reduction	3/6 (50)	5/6 (83)
0%–30% reduction	3/8 (37)	2/8 (25)
0%–20% increase	2/9 (22)	1/9 (11)
>20% increase	0/3 (0)	1/3 (33)
Unevaluable SLD	1/1 (100)	1/1 (100)
Objective response per RECIST v1.1		
Best response of PR	3/5 (60)	4/5 (80)
Best response of SD	5/12 (42)	2/12 (17)
Best response of PD	1/11 (9)	4/11 (36)

b

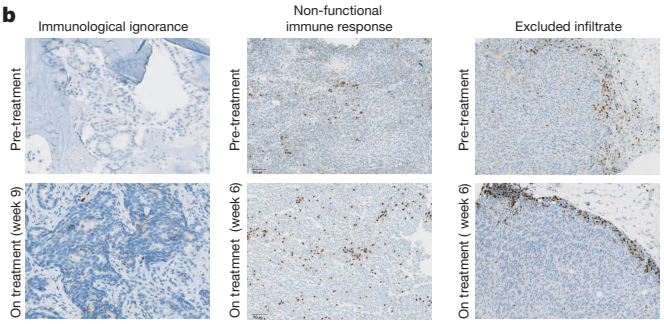


Figure 4 | Biomarker status and response to MPDL3280A. a, Table summarizing the frequency of patients with an increase in PD-L1-positive tumour-infiltrating immune cells (ICs) and tumour cells (TCs) by change in the sum of the longest diameter (SLD) and by response to MPDL3280A in patients with paired serial biopsies. There were 28 paired serial biopsies. Of these, 16 tumours were melanoma, 4 were renal cell carcinoma, 4 were non-small cell lung cancer, 2 were head and neck carcinoma, and 2 were colorectal cancer. Patients with an increase of ≥5% in PD-L1-expressing tumour cells and tumour-infiltrating immune cells were identified as having increased PD-L1 expression by IHC after treatment with MPDL3280A. The patient who was unevaluable for SLD had the responding tumour excised for biomarker analysis. This table also includes one patient with progressive disease (PD) by RECIST version 1.1 but without post-dose SLD measures. **b**, Left: 'immunological ignorance' visualized by CD8 IHC. See Extended Data Fig. 8a for additional information. Middle: 'non-functional immune response' visualized by CD8 IHC. See Extended Data Fig. 8b for additional information. Right: 'excluded infiltrate' visualized by CD8 IHC. See Extended Data Fig. 9 for additional information. All images are at 10× magnification. PR, partial response; RECIST, Response Evaluation Criteria in Solid Tumours; SD, stable disease.

In recent years it has become clear that modulation of a patient's immune system can be an effective cancer therapy[19–22]; however, our understanding of human cancer immunology is incomplete. Therefore, as part of our phase Ia dose escalation and expansion study with MPDL3280A, we focused on understanding the biomarkers relating to the PD-L1–PD-1 pathway.

This MPDL3280A study did not follow the design of a traditional phase I clinical trial, but instead enrolled large numbers of patients with defined characteristics into expansion phases. MPDL3280A doses ranged from 0.01 to 20 mg kg^{-1} q3w and clinical activity was seen from 1 to 20 mg kg^{-1}. The maximum tolerated dose of MPDL3280A was not reached, and no dose-limiting toxicities were observed (ref. 18). Because 15 mg kg^{-1} q3w was sufficient to maintain target drug levels (based on clinical and non-clinical information), the equivalent fixed dose of 1,200 mg q3w is being moved forward in clinical development as monotherapy. The accompanying report (ref. 18) additionally describes the activity of MPDL3280A in bladder cancer.

In addition to observing clinical responses greater than the historic averages in these refractory patient populations, our most important finding was the association of PD-L1 expression with clinical response to MPDL3280A. It was unexpected that the association of tumour-infiltrating immune cell PD-L1 expression with treatment response appeared stronger than that with tumour cell PD-L1 expression. This finding appears inconsistent with a simple 'adaptive response' hypothesis where T-cell-derived IFN-γ induces protective expression of PD-L1 by the tumour cells[12,23]. While upregulation of PD-L1 by tumour cells occurred

post-treatment in responders, our results instead suggest that tumour-infiltrating immune cells may be more sensitive to IFN-γ expression and may act preferentially to suppress pre-existing T-cell responses before therapy. These data indicate that, although additional immune regulatory pathways may be involved, PD-L1 appears to have a dominant role in direct T-cell immunosuppression. Furthermore, intratumoral expression of PD-L2 did not affect the response to anti-PD-L1, and the expression of other T-cell negative regulators also failed to correlate with poor response. Although we cannot exclude the possibility that inhibiting these receptors might enhance responses to PD-L1 blockade, it was striking that MPDL3280A was effective despite their presence[24,25].

Higher pre-treatment expression of CTLA4 was observed to correlate with response to MPDL3280A. These results suggest that CTLA4, although an important regulator during T-cell expansion, is also a marker of the presence of activated T cells whose functional role as a negative regulator of intra-tumoral T cells appears to be less important than that of PD-L1 (refs 19, 22). Another correlation was that of elevated pre-treatment fractalkine expression with disease progression. This result was unexpected because this chemokine is generally associated with T-cell infiltration.

We also examined blood-based biomarkers. The observed rise in ITAC—an IFN-γ inducible chemokine that is chemotactic for activated T cells[26] and IL-18, a pro-inflammatory cytokine whose presence generally induces, rather than reflects, IFN-γ release—suggests the rapid expansion of a pre-existing primed immune state, perhaps even extra-tumorally. The increases in activated cytotoxic T lymphocytes during this same time frame and the clinical reports of fever during cycle 1 further support this notion[27,28] and indicate that PD-L1 blockade may also contribute to an overall expansion of the T-cell compartment at the level of antigen-presenting cells. The decrease in IL-6 may be indicative of the opposing role of effector T cells and suppressive myeloid cells. Given that these changes do not clearly segregate responding patients, they may reflect a systemic re-priming and expansion of both pre-existing anti-tumour T-cell and non-tumour-directed T-cell populations.

In summary, this study analysed the mechanisms associated with clinical response and lack of response to MPDL3280A, providing evidence for the 'inflamed tumour' hypothesis[12,29]. However, larger studies will be needed to study the relationship between PD-L1 expression and patient survival. Pre-existing immunity is probably necessary for most responses, and is further amplified during treatment. While important to further characterize the immune profile of responders, understanding the profile of non-responders will probably provide even more valuable information, possibly revealing the diversity of mechanisms controlling antitumour immunity and suggesting new strategies to promote the cancer immunity cycle[5].

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 May; accepted 31 October 2014.

1. Mellman, I., Coukos, G. & Dranoff, G. Cancer immunotherapy comes of age. *Nature* **480,** 480–489 (2011).
2. Chen, D. S., Irving, B. A. & Hodi, F. S. Molecular pathways: next-generation immunotherapy–inhibiting programmed death-ligand 1 and programmed death-1. *Clin. Cancer Res.* **18,** 6580–6587 (2012).
3. van Rooij, N. *et al.* Tumor exome analysis reveals neoantigen-specific T-cell reactivity in an ipilimumab-responsive melanoma. *J. Clin. Oncol.* **31,** e439–e442 (2013).
4. Gros, A. *et al.* PD-1 identifies the patient-specific $CD8^+$ tumor-reactive repertoire infiltrating human tumors. *J. Clin. Invest.* **124,** 2246–2259 (2014).
5. Chen, D. S. & Mellman, I. Oncology meets immunology: the cancer-immunity cycle. *Immunity* **39,** 1–10 (2013).
6. Park, J. J. *et al.* B7-H1/CD80 interaction is required for the induction and maintenance of peripheral T-cell tolerance. *Blood* **116,** 1291–1298 (2010).
7. Yang, J. *et al.* The novel costimulatory programmed death ligand 1/B7.1 pathway is functional in inhibiting alloimmune responses *in vivo*. *J. Immunol.* **187,** 1113–1119 (2011).
8. Paterson, A. M. *et al.* The programmed death-1 ligand 1:B7-1 pathway restrains diabetogenic effector T cells *in vivo*. *J. Immunol.* **187,** 1097–1105 (2011).
9. Butte, M. J., Keir, M. E., Phamduy, T. B., Sharpe, A. H. & Freeman, G. J. Programmed death-1 ligand 1 interacts specifically with the B7-1 costimulatory molecule to inhibit T cell responses. *Immunity* **27,** 111–122 (2007).
10. Dong, H., Zhu, G., Tamada, K. & Chen, L. B7-H1, a third member of the B7 family, co-stimulates T-cell proliferation and interleukin-10 secretion. *Nature Med.* **5,** 1365–1369 (1999).
11. Day, C. L. *et al.* PD-1 expression on HIV-specific T cells is associated with T-cell exhaustion and disease progression. *Nature* **443,** 350–354 (2006).
12. Taube, J. M. *et al.* Colocalization of inflammatory response with B7-h1 expression in human melanocytic lesions supports an adaptive resistance mechanism of immune escape. *Sci. Transl. Med.* **4,** 127ra37 (2012).
13. Matsumoto, K. *et al.* B7-DC induced by IL-13 works as a feedback regulator in the effector phase of allergic asthma. *Biochem. Biophys. Res. Commun.* **365,** 170–175 (2008).
14. Akbari, O. *et al.* PD-L1 and PD-L2 modulate airway inflammation and iNKT-cell-dependent airway hyperreactivity in opposing directions. *Mucosal Immunol.* **3,** 81–91 (2010).
15. Isaacs, J. D. *et al.* A therapeutic human IgG4 monoclonal antibody that depletes target cells in humans. *Clin. Exp. Immunol.* **106,** 427–433 (1996).
16. Warncke, M. *et al.* Different adaptations of IgG effector function in human and nonhuman primates and implications for therapeutic antibody treatment. *J. Immunol.* **188,** 4405–4411 (2012).
17. Wolchok, J. D. *et al.* Guidelines for the evaluation of immune therapy activity in solid tumors: immune-related response criteria. *Clin. Cancer Res.* **15,** 7412–7420 (2009).
18. Powles, T. *et al.* MPDL3280A treatment leads to clinical activity in metastatic bladder cancer. *Nature* http://dx.doi.org/10.1038/nature13904 (2014).
19. Hodi, F. S. *et al.* Improved survival with ipilimumab in patients with metastatic melanoma. *N. Engl. J. Med.* **363,** 711–723 (2010).
20. Brahmer, J. R. *et al.* Safety and activity of anti-PD-L1 antibody in patients with advanced cancer. *N. Engl. J. Med.* **366,** 2455–2465 (2012).
21. Topalian, S. L. *et al.* Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *N. Engl. J. Med.* **366,** 2443–2454 (2012).
22. Hamid, O. *et al.* Safety and tumor responses with lambrolizumab (anti-PD-1) in melanoma. *N. Engl. J. Med.* **369,** 134–144 (2013).
23. Dong, H. *et al.* Tumor-associated B7–H1 promotes T-cell apoptosis: a potential mechanism of immune evasion. *Nature Med.* **8,** 793–800 (2002).
24. Park, H. J. *et al.* Tumor-infiltrating regulatory T cells delineated by upregulation of PD-1 and inhibitory receptors. *Cell. Immunol.* **278,** 76–83 (2012).
25. Spranger, S. *et al.* Up-regulation of PD-L1, IDO, and T_{regs} in the melanoma tumor microenvironment is driven by $CD8^+$ T cells. *Sci. Transl. Med.* **5,** 200ra116 (2013).
26. Cole, K. E. *et al.* Interferon-inducible T cell alpha chemoattractant (I-TAC): a novel non-ELR CXC chemokine with potent activity on activated T cells through selective high affinity binding to CXCR3. *J. Exp. Med.* **187,** 2009–2021 (1998).
27. Wrammert, J. *et al.* Rapid cloning of high-affinity human monoclonal antibodies against influenza virus. *Nature* **453,** 667–671 (2008).
28. Han, A. *et al.* Dietary gluten triggers concomitant activation of $CD4^+$ and $CD8^+$ αβ T cells and γδ T cells in celiac disease. *Proc. Natl Acad. Sci. USA* **110,** 13073–13078 (2013).
29. Gajewski, T. F., Schreiber, H. & Fu, Y. X. Innate and adaptive immune cells in the tumor microenvironment. *Nature Immunol.* **14,** 1014–1022 (2013).

Acknowledgements We thank the patients and their families. We also thank all of the investigators and their staff, including A. Balmanoukian and P. Boasberg from The Angeles Clinic and Research Institute; T. Powles from Barts Cancer Institute, QMUL, Barts Health NHS Trust; D. Cho from NYU Langone Medical Center; P. Cassier from Centre Léon-Bérard; F. Braiteh from USON Research Network, Comprehensive Cancer Centers of Nevada; N. Vogelzang from USON Research Network, Comprehensive Cancer Centers of Nevada and University of Nevada; T. Choueiri, L. Gandhi, N. Ibrahim and P. Ott from Dana-Farber Cancer Institute; J.-P. Delord and C. Gomez-Rocca from Institut Claudius Regaud; A. Hollebecque and R. Bahleda from Gustave Roussy; L. Emens from Johns Hopkins Medicine, The Sidney Kimmel Comprehensive Cancer Center; K. Flaherty and R. Sullivan from Massachusetts General Hospital; S. Antonia from Moffitt Cancer Center; H. Burris, J. Infante and D. Spigel from Sarah Cannon Research Institute; G. Fisher from Stanford Medicine, Cancer Institute; P. Conkling and L. Garbo from US Oncology Research, Inc.; C. Cruz and J. Tabenero from Vall d'Hebron Institute of Oncology and Vall d'Hebron University Hospital; W. Pao and I. Puzanov from Vanderbilt-Ingram Cancer Center; P. Eder, H. Kluger and M. Sznol from Yale Cancer Center. From Genentech, we thank M. Anderson, M. Boe, Z. Boyd, C. Chappey, M. Denker, R. Desai, L. Fu, B. Irving, D. Jin, W. Kadel, R. Nakamura, I. Rhee, X. Shen, M. Stroh, T. Sumiyoshi, J. Wu, Y. Xin and J. Yi. Support for third-party writing assistance for this manuscript was provided by F. Hoffmann-La Roche Ltd. NCI grants 1R01CA155196 (Battle-2) and P30 CA 016359 (CCSG) to R.S.H. helped support the infrastructure for this trial and program.

Author Contributions R.S.H., J.-C.S., D.S.C., F.S.H. and J.A.S. contributed to the overall study design. M.K., S.R., Y.X., H.K. and P.S.H. provided the biomarker studies. M.L. performed the pharmacokinetic analysis. I.M. provided the pre-clinical analysis. A.M. performed the statistical analysis. All authors analysed the data. All authors contributed to writing the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.S.H. (roy.herbst@yale.edu).

LETTER

doi:10.1038/nature13954

PD-1 blockade induces responses by inhibiting adaptive immune resistance

Paul C. Tumeh[1,2], Christina L. Harview[1], Jennifer H. Yearley[3], I. Peter Shintaku[1], Emma J. M. Taylor[1], Lidia Robert[1], Bartosz Chmielowski[1,2], Marko Spasic[1], Gina Henry[1], Voicu Ciobanu[1], Alisha N. West[1], Manuel Carmona[1], Christine Kivork[1], Elizabeth Seja[1], Grace Cherry[1], Antonio J. Gutierrez[1], Tristan R. Grogan[1], Christine Mateus[4], Gorana Tomasic[4], John A. Glaspy[1,2], Ryan O. Emerson[5], Harlan Robins[5,6], Robert H. Pierce[3], David A. Elashoff[1,2], Caroline Robert[4] & Antoni Ribas[1,2]

Therapies that target the programmed death-1 (PD-1) receptor have shown unprecedented rates of durable clinical responses in patients with various cancer types[1–5]. One mechanism by which cancer tissues limit the host immune response is via upregulation of PD-1 ligand (PD-L1) and its ligation to PD-1 on antigen-specific CD8$^+$ T cells (termed adaptive immune resistance)[6,7]. Here we show that pre-existing CD8$^+$ T cells distinctly located at the invasive tumour margin are associated with expression of the PD-1/PD-L1 immune inhibitory axis and may predict response to therapy. We analysed samples from 46 patients with metastatic melanoma obtained before and during anti-PD-1 therapy (pembrolizumab) using quantitative immunohistochemistry, quantitative multiplex immunofluorescence, and next-generation sequencing for T-cell antigen receptors (TCRs). In serially sampled tumours, patients responding to treatment showed proliferation of intratumoral CD8$^+$ T cells that directly correlated with radiographic reduction in tumour size. Pre-treatment samples obtained from responding patients showed higher numbers of CD8-, PD-1- and PD-L1-expressing cells at the invasive tumour margin and inside tumours, with close proximity between PD-1 and PD-L1, and a more clonal TCR repertoire. Using multivariate analysis, we established a predictive model based on CD8 expression at the invasive margin and validated the model in an independent cohort of 15 patients. Our findings indicate that tumour regression after therapeutic PD-1 blockade requires pre-existing CD8$^+$ T cells that are negatively regulated by PD-1/PD-L1-mediated adaptive immune resistance.

Recently, we reported sustained tumour regression in 38% of patients in a multi-institutional, international, phase 1 expansion study evaluating the safety and clinical activity of pembrolizumab (formerly known as MK-3475 and lambrolizumab), a humanized monoclonal antibody against PD-1, in patients with advanced melanoma (http://ClinicalTrials.gov study number NCT01295827)[3,8]. PD-L1, known to be expressed by cells in the tumour microenvironment, engages PD-1 on T cells and subsequently triggers inhibitory signalling downstream of the TCR, blocking effector functions and reducing T-cell killing capacity[6]. PD-L1 can be constitutively expressed on the surface of cancer cells through poorly characterized oncogenic signalling pathways[9,10], or alternatively, expressed in response to the presence of T cells producing immune-stimulating cytokines such as interferons[7,11,12]. The process of expression of PD-L1 in response to cytokines has been termed adaptive immune resistance[6], and represents a mechanism by which cancer cells attempt to protect themselves from immune-cell-mediated killing.

We sought to determine whether pre-existing tumour-associated CD8$^+$ T cells inhibited by PD-1/PD-L1 engagement represent key factors in determining clinical response to PD-1 blocking therapy. Our study cohort consisted of 46 patients with advanced melanoma treated with single-agent pembrolizumab between December 2011 and October 2013 at UCLA (Institutional Review Board (IRB) study number 11-003066). Patients underwent tumour biopsies before and during treatment. Baseline biopsy samples from 15 additional patients with advanced melanoma enrolled in the same pembrolizumab phase 1 clinical trial at Gustave Roussy in Villejuif, Paris, France (IRB study number 11-040) were analysed as a validation cohort (Extended Data Table 1).

We first examined the spatiotemporal dynamics of CD8$^+$ T cells by performing qualitative and quantitative immunohistochemistry (IHC) analysis for CD8 expression before and during PD-1 blockade in two tumour compartments: the invasive tumour margin (stromal–tumour edge) and inside the tumour parenchyma (tumour centre)[13,14]. S100 expression was used to define the invasive margin and tumour centre (Extended Data Fig. 1a). Pre-treatment samples obtained from patients who experienced a tumour response (response group; Fig. 1a), showed higher CD8$^+$-cell densities at the invasive margin when compared to samples from patients who progressed during therapy (progression group; Fig. 1b). Extended Data Table 2 provides the anatomical location of all tumours serially sampled. Serially sampled tumours during treatment exhibited a parallel increase in CD8$^+$-cell density at both the invasive margin and tumour centre in the response group (Spearman's correlation $r = 0.71$, $P < 0.001$; Fig. 1c), but not in the progression group (Fig. 1d). Two patients experienced delayed responses (Fig. 1c, triangles) and showed step-wise accumulation of CD8$^+$ cells, with initial increases restricted to the invasive margin, followed by mobilization into the tumour parenchyma (Extended Data Fig. 1b).

Releasing the PD-1 immune checkpoint in pre-existing tumour-antigen-specific T cells should lead to T-cell proliferation, intratumoral infiltration and increased effector function. We found a greater increase in CD8$^+$ density from baseline to post-dosing biopsy that significantly correlated with a decrease in radiographic tumour size (Extended Data Fig. 2a; Spearman's correlation $r = -0.75$, $P = 0.0002$). During treatment, we found an increase in cells that were double positive for CD8 and the nuclear proliferation marker Ki67 in samples from patients with a tumour response. We observed all sub-phases of mitosis based on characteristic chromatin patterns (Fig. 2 and Extended Data Fig. 2b, c). The post-dosing increase in CD8/Ki67 double-positive cells was restricted to the tumour parenchyma. We found increased expression of granzyme B, a cytotoxic granule reflective of CD8 effector function, on CD8$^+$ cells in post-dosing biopsies in the response group ($P < 0.0001$; Extended Data Fig. 3a, b).

The correlation between T-cell activation/effector function and treatment outcome upon release of the PD-1 immune checkpoint may be driven by the production of interferons by tumour-infiltrating CD8 cells that induce PD-L1 expression on tumour-resident cells[11,15]. To test this hypothesis, we stained baseline and post-dosing biopsies for phospho-STAT1 (pSTAT1), which is an immediate downstream effector upon interferon-γ binding to its receptor (Extended Data Fig. 3c, d). The response group was associated with significantly higher expression of pSTAT1 at the invasive margin, localized to the area of CD8 infiltrate, before ($P = 0.002$) and during treatment ($P < 0.0001$), when compared to biopsies from the progression group. In serially sampled tumours from

[1]University of California Los Angeles (UCLA), Los Angeles, California 90095, USA. [2]Jonsson Comprehensive Cancer Center, Los Angeles, California 90095, USA. [3]Merck & Co, Palo Alto, California 94304, USA. [4]Gustave Roussy and INSERM U981, Villejuif, Paris Sud, France. [5]Adaptive Biotechnologies, Seattle, Washington 98102, USA. [6]Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA.

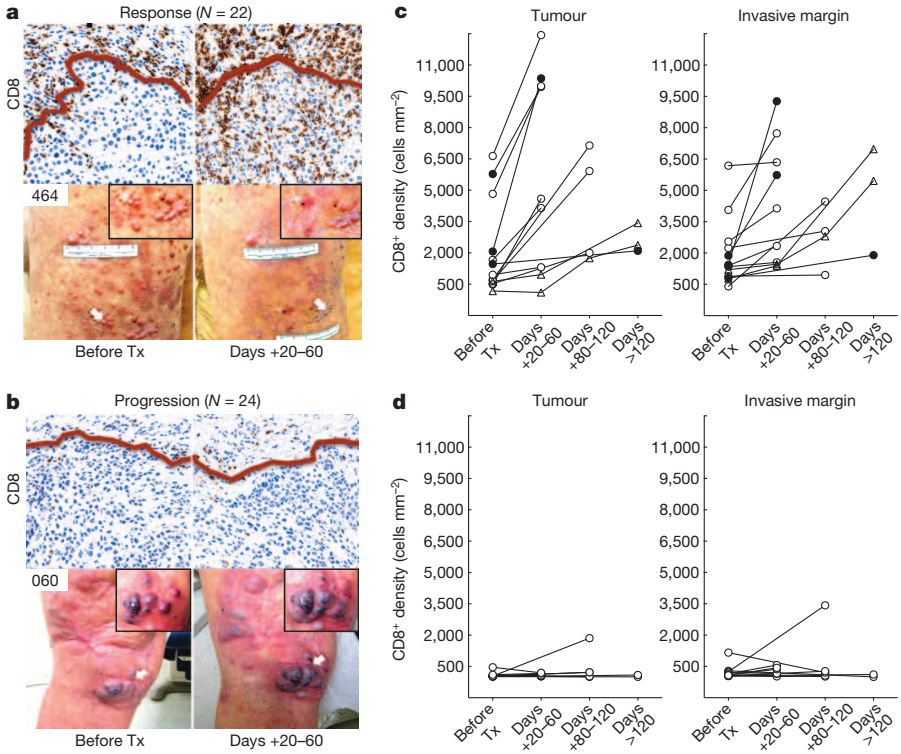


Figure 1 | Immunohistochemical analysis of CD8$^+$ T cells in samples obtained before and during pembrolizumab treatment. **a**, **b**, Examples of CD8 expression in melanoma tumours serially biopsied before PD-1 blocking treatment (Tx) and 20–60 days after treatment began (Days +20–60) from a patient in the response (**a**) and progression (**b**) groups. Red line separates tumour parenchyma (below line) and invasive margin (above line). Magnification, ×20. **c**, **d**, CD8$^+$-cell density at the tumour parenchyma and invasive margin in samples from all responders (**c**; $n = 13$) and progressors (**d**; $n = 12$) who received a biopsy before and during treatment. Filled circle indicates complete response; open circle indicates partial response; triangle indicates delayed response.

the response group, pSTAT1 expression was also found to be significantly higher during treatment when compared to baseline ($P = 0.007$) (Extended Data Fig. 3e). These findings prompted us to investigate the association between CD8$^+$-, CD4$^+$-, PD-1$^+$- and PD-L1$^+$-cell densities in baseline biopsies in regards to response to treatment (Fig. 3a). The response group was associated with significantly higher numbers of CD8$^+$, PD-1$^+$ and PD-L1$^+$ cells at both the invasive margin and the tumour centre when compared to the progression group (CD8, $P < 0.0001$; PD-1, $P = 0.0002$; PD-L1, $P = 0.006$). However, CD4 expression at baseline was not found to correlate with treatment outcome. No relationship was found between previous treatment history with ipilimumab (anti-CTLA4) and CD8 expression or treatment outcome (Extended Data Table 3).

We next determined the relative proximity of PD-1 and PD-L1 as evidence of a physical interaction between PD-1$^+$ and PD-L1$^+$ cells, a presumptive requisite for adaptive immune resistance. Figure 3b shows representative examples of chromogenic PD-1 and PD-L1 expression in serially cut tissue sections as well as multiplexed PD-1 × PD-L1 immunofluorescence in pre-treatment samples according to treatment outcome. Using quantitative multiplexed PD-1 × PD-L1 immunofluorescence, we found a significant correlation between proximity of PD-1 and PD-L1 and response to therapy (Fig. 3c; $P = 0.005$).

We next investigated the relationship between CD8 and PD-L1 using a Spearman's correlation analysis and found the two markers to correlate in both the tumour (Spearman $r = 0.598$, $P < 0.001$) and the invasive

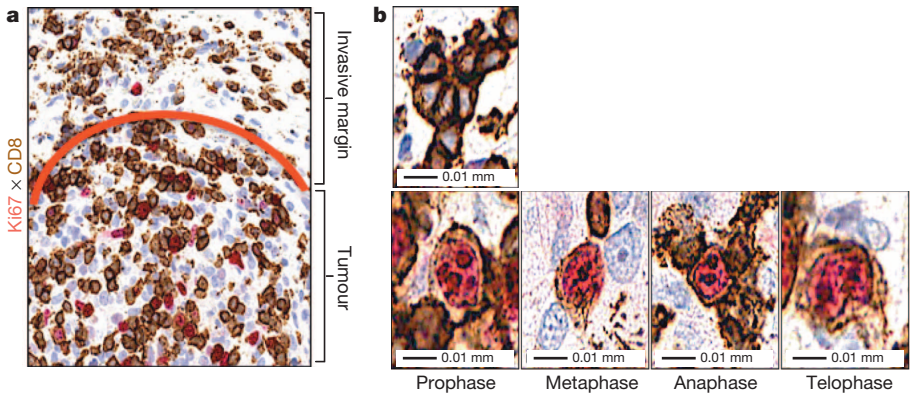


Figure 2 | Regressing tumours during treatment are associated with proliferating CD8$^+$ T cells that localize to the tumour. **a**, Representative example of CD8/Ki67 chromogenic double staining from a sample obtained during tumour regression shows double-positive T cells localized to the tumour parenchyma. The red line separates the invasive margin (above line) and tumour (below line). **b**, Top, representative single-positive quiescent CD8$^+$ brown cells (no Ki67 labelling) from the invasive margin. Bottom, representative double-positive cells (red, labelled Ki67 nucleus; brown, labelled CD8 membrane) with characteristic chromatin patterns associated with sub-phases of mitosis. Magnification, ×40.

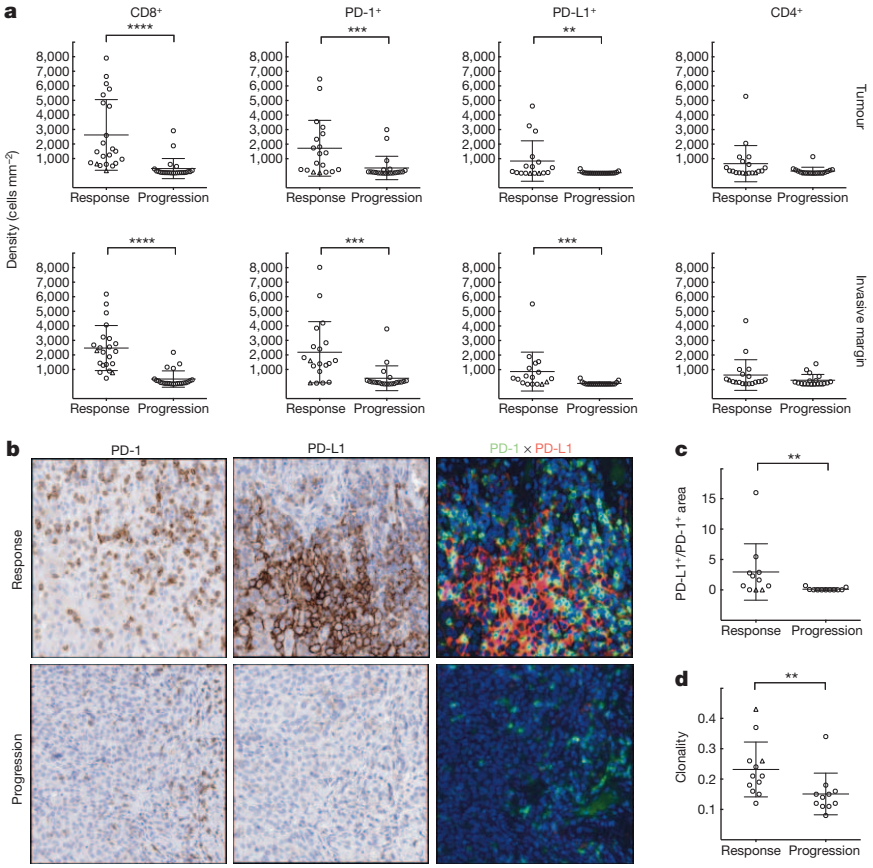


Figure 3 | Baseline density, location and proximity of CD8$^+$, PD-1$^+$, PD-L1$^+$ and CD4$^+$ cells, and T-cell repertoire according to treatment outcome. **a**, Melanoma samples collected before treatment with PD-1 blocking therapy were assessed for CD8 (response $n = 22$, progression $n = 24$), PD-1 (response $n = 19$, progression $n = 21$), PD-L1 (response $n = 17$, progression $n = 21$) and CD4 (response $n = 19$, progression $n = 18$) density by quantitative immunohistochemistry in the tumour compartment and at the invasive margin. $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$. **b**, Examples of the relative proximity of PD-1- and PD-L1-expressing cells in representative baseline samples from a responder and a progressor. **c**, Proximity analysis of PD-1 and PD-L1 based on multiplex quantitative immunofluorescence in baseline tumour samples (response $n = 11$, progression $n = 11$). $^{**}P = 0.005$. **d**, Results of TCR sequencing performed on 25 whole tumour samples taken at baseline (response $n = 12$, progression $n = 11$). Triangles indicate delayed response. $^{**}P = 0.004$. Error bars denote the s.d.; horizontal lines denote the mean.

margin (Spearman $r = 0.527$, $P < 0.001$). Furthermore, CD8 and PD-L1 density co-varied with treatment outcome in both the tumour and invasive margin ($P < 0.001$ for both; Extended Data Fig. 4a, b). Immunofluorescence multiplexing for CD8 and PD-1 corroborated our chromogenic IHC findings that CD8$^+$ T cells represented the primary cellular source of PD-1 expression (Extended Data Fig. 4c). Using chromogenic double staining for SOX10 and PD-L1, we found PD-L1 expression on melanoma cells and also on cells morphologically consistent with lymphocytes and macrophages in samples obtained during treatment from the response group (Extended Data Fig. 5a). Principal component analysis of samples obtained before treatment showed that CD8, PD-1 and PD-L1 expression in the tumour ($P = 0.001$) and at the invasive margin significantly correlated with treatment outcome ($P < 0.0001$; Extended Data Fig. 5b).

The high density of CD8$^+$ cells at the site of the tumour in the response group is suggestive of a specific immune response to tumour antigens. Therefore, we hypothesized that a more restricted TCR sequence usage would reflect a tumour-antigen-specific T-cell accumulation at the tumour site. Using genomic DNA isolated from pre-treatment samples, we performed next-generation sequencing to capture all uniquely rearranged, variable TCR β-chain regions[16,17]. We found that a more restricted TCR β-chain usage, reflecting a T-cell population that was less diverse in repertoire and more clonal in nature, significantly correlated with clinical response to pembrolizumab treatment ($P = 0.004$; Fig. 3d). The clonality read-out was not found to highly correlate with tumour-infiltrating lymphocyte density ($R^2 = 0.04$; Extended Data Fig. 6a). However, biopsies from patients with a tumour response showed evidence of an enriched population of T cells with unique specificities. In addition, comparison of the TCR clonality at baseline and post-dosing biopsies showed that in samples from the response group, more than ten times as many clones expanded after anti-PD-1 therapy than in the progression group (Extended Data Fig. 6b, c).

To create the best discriminatory model to assess the probability of clinical response to PD-1 blocking therapies, forward stepwise logistic regression was run on CD8$^+$-, CD4$^+$-, PD-1$^+$- and PD-L1$^+$-cell densities within the tumour and the invasive margin. Results of the stepwise procedure, and a logistic regression model, consistently selected the invasive margin CD8$^+$ density as the best full predictive parameter (Extended Data Table 4a). The next best predictors were tumour CD8$^+$ T-cell density, tumour and invasive margin PD-1$^+$ density, and tumour and invasive margin PD-L1$^+$ density. Tumour and invasive margin CD4$^+$ density were the poorest predictors.

To test this predictive model, we obtained pre-treatment biopsies from 15 patients treated at Gustave Roussy and were blinded to treatment outcome. We quantified CD8$^+$ T-cell density in the invasive margin and used our logistic model to calculate a predicted probability of response for each patient in the validation cohort (Extended Data Table 4b). Out of the 15 patients, we accurately predicted 4 out of 5 patients in the true progression group and 9 out of 9 patients in the true response group. There was one false-positive prediction and one patient predicted to respond who remains in stable disease.

Our studies build upon the evidence that response rates to PD-1 or PD-L1 blocking antibodies are higher in patients whose tumours express

PD-L1 (refs 1, 15). Since PD-L1 can be either constitutively expressed or induced upon T-cell recognition and production of interferons[11,15], we hypothesized that response to PD-1 blockade would more tightly covariate with the inducible PD-L1 expression in the presence of antigen-specific T cells[7], termed adaptive immune resistance[6]. Indeed, we found PD-1-positive T cells interfacing with PD-L1-expressing cells within tumours in pre-treatment samples from responders. The clinical relevance of the relative distribution of PD-L1 expression on cancer cells, myeloid-derived cells and activated T cells in tumours, in terms of treatment outcome, remains to be elucidated. Our data suggest that PD-L1 may serve as an indirect marker of adaptive immune resistance in response to tumour-antigen-specific T-cell infiltration rather than as a static constitutive biomarker. Hence, inducing a type-I interferon inflammatory response in combination with PD-L1 blockade merits further clinical investigation[11].

T-cell infiltrates have been found to have predictive value with respect to the natural history of primary cancers[13,14,18]. We build on this and report that the baseline density and location of T cells in metastatic melanomas have predictive value in the treatment outcome of patients receiving therapies that block the PD-1/PD-L1 axis. Releasing the PD-1 immune checkpoint results in clinically relevant anti-tumour activity when there is a greater density of pre-existing tumour-antigen-restricted $CD8^+$ T cells that are negatively regulated by PD-1/PD-L1 interactions.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 20 May; accepted 9 October 2014.

1. Topalian, S. L. *et al.* Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *N. Engl. J. Med.* **366,** 2443–2454 (2012).
2. Brahmer, J. R. *et al.* Safety and activity of anti-PD-L1 antibody in patients with advanced cancer. *N. Engl. J. Med.* **366,** 2455–2465 (2012).
3. Hamid, O. *et al.* Safety and tumor responses with lambrolizumab (anti-PD-1) in melanoma. *N. Engl. J. Med.* **369,** 134–144 (2013).
4. Wolchok, J. D. *et al.* Nivolumab plus ipilimumab in advanced melanoma. *N. Engl. J. Med.* **369,** 122–133 (2013).
5. Topalian, S. L. *et al.* Survival, durable tumor remission, and long-term safety in patients with advanced melanoma receiving nivolumab. *J. Clin. Oncol.* **32,** 1020–1030 (2014).
6. Pardoll, D. M. The blockade of immune checkpoints in cancer immunotherapy. *Nature Rev. Cancer* **12,** 252–264 (2012).
7. Spranger, S. *et al.* Up-regulation of PD-L1, IDO, and T_{regs} in the melanoma tumor microenvironment is driven by $CD8^+$ T cells. *Sci. Transl. Med.* **5,** 200ra116 (2013).
8. Robert, C. *et al.* Anti-programmed-death-receptor-1 treatment with pembrolizumab in ipilimumab-refractory advanced melanoma: a randomised dose-comparison cohort of a phase 1 trial. *Lancet* **384,** 1109–1117 (2014).
9. Parsa, A. T. *et al.* Loss of tumor suppressor PTEN function increases B7-H1 expression and immunoresistance in glioma. *Nature Med.* **13,** 84–88 (2007).
10. Atefi, M. *et al.* Effects of MAPK and PI3K pathways on PD-L1 expression in melanoma. *Clin. Cancer Res.* **20,** 3446–3457 (2014).
11. Bald, T. *et al.* Immune cell-poor melanomas benefit from PD-1 blockade after targeted type I IFN activation. *Cancer Discov* **4,** 674–687 (2014).
12. Duraiswamy, J., Freeman, G. J. & Coukos, G. Dual blockade of PD-1 and CTLA-4 combined with tumor vaccine effectively restores T-cell rejection function in tumors–response. *Cancer Res.* **74,** 633–634 (2014).
13. Galon, J. *et al.* Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science* **313,** 1960–1964 (2006).
14. Pagès, F. *et al.* Effector memory T cells, early metastasis, and survival in colorectal cancer. *N. Engl. J. Med.* **353,** 2654–2666 (2005).
15. Taube, J. M. *et al.* Association of PD-1, PD-1 ligands, and other features of the tumor immune microenvironment with response to anti-PD-1 therapy. *Clin. Cancer Res.* **20,** 5064–5074 (2014).
16. Robins, H. S. *et al.* Comprehensive assessment of T-cell receptor β-chain diversity in αβ T cells. *Blood* **114,** 4099–4107 (2009).
17. Carlson, C. S. *et al.* Using synthetic templates to design an unbiased multiplex PCR assay. *Nat. Commun.* **4,** 2680 (2013).
18. Zhang, L. *et al.* Intratumoral T cells, recurrence, and survival in epithelial ovarian cancer. *N. Engl. J. Med.* **348,** 203–213 (2003).

Acknowledgements This study was funded in part by National Institutes of Health grants K08 AI091663, Kure It Research Grant, UL1TR000124 (to P.C.T.), P01 CA168585, U54 CA119347, R01 CA170689, the Ressler Family Fund, the Dr Robert Vigen Memorial Fund, the Wesley Coyle Memorial Fund, and the Garcia-Corsini Family Fund (to A.R.), P30 CA16042 to D.A.E. A.R. was supported by a Stand Up To Cancer—Cancer Research Institute Cancer Immunology Dream Team Translational Research Grant (SU2C-AACR-DT1012). Stand Up To Cancer is a program of the Entertainment Industry Foundation administered by the American Association for Cancer Research. M.S. was supported as a Howard Hughes Medical Institute Medical Research Fellow. Some of the studies were funded by Merck Sharp and Dome and by Adaptive Biotechnologies. L.R. was supported by the V Foundation-Gil Nickel Family Endowed Fellowship in Melanoma Research and a grant from the Spanish Society of Medical Oncology for Translational Research in Reference Centers. We acknowledge the Translational Pathology Core Laboratory for tissue sectioning and slide scanning, S. Roy, N. Kamsu-Kom, R. Guo, J. Pang, W. Li, A. Villanueva and K. Crawford for biopsy processing and clinical data, S. Hashaghan for assisting with quantitative imaging approaches, E. Penaflor for assisting in IHC assay optimization, execution and digital image generation, and B. Dogdas and S. Mehta who assisted with the proximity assay. We would like to thank S. Ebbinghaus, E. Rubin, S. P. Kang, R. L. Modlin, C. R. Taylor and C. Denny for critically reviewing the manuscript.

Author Contributions P.C.T. and A.R. supervised the project and developed the concepts. P.C.T., C.L.H., C.R. and A.R. designed the experiments. P.C.T., C.L.H., M.S., C.R. and A.R. interpreted the data. A.R., P.C.T., C.L.H., C.R., J.H.Y., M.S., I.P.S., E.J.M.T., R.O.E., H.R. and R.H.P. gave conceptual advice and edited the manuscript. P.C.T., J.H.Y., I.P.S., A.N.W. and E.J.M.T. established IHC staining and/or imaging protocols. J.H.Y., I.P.S., E.J.M.T. and R.H.P. provided confirmatory pathology analyses. G.T. worked on IHC samples from the patients from Gustave Roussy. C.L.H., L.R., M.S., G.H., V.C., M.C., C.K. and E.S. provided technical support. P.C.T., T.R.G. and D.A.E. designed and implemented the predictive model and provided statistical support. P.C.T., B.C., A.J.G., C.M., J.A.G., G.C., C.R. and A.R. clinically evaluated patients in the trial.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Readers are welcome to comment on the online version of the paper. The authors declare competing financial interests: details are available in the online version of the paper. Correspondence and requests for materials should be addressed to P.C.T. (ptumeh@mednet.ucla.edu) or A.R. (aribas@mednet.ucla.edu).

LETTER

doi:10.1038/nature14001

Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing

Mahesh Yadav[1]*, Suchit Jhunjhunwala[1]*, Qui T. Phung[1], Patrick Lupardus[1], Joshua Tanguay[1], Stephanie Bumbaca[1], Christian Franci[1], Tommy K. Cheung[1], Jens Fritsche[2], Toni Weinschenk[2], Zora Modrusan[1], Ira Mellman[1], Jennie R. Lill[1]§ & Lélia Delamarre[1]§

Human tumours typically harbour a remarkable number of somatic mutations[1]. If presented on major histocompatibility complex class I molecules (MHCI), peptides containing these mutations could potentially be immunogenic as they should be recognized as 'non-self' neo-antigens by the adaptive immune system. Recent work has confirmed that mutant peptides can serve as T-cell epitopes[2–9]. However, few mutant epitopes have been described because their discovery required the laborious screening of patient tumour-infiltrating lymphocytes for their ability to recognize antigen libraries constructed following tumour exome sequencing. We sought to simplify the discovery of immunogenic mutant peptides by characterizing their general properties. We developed an approach that combines whole-exome and transcriptome sequencing analysis with mass spectrometry to identify neo-epitopes in two widely used murine tumour models. Of the >1,300 amino acid changes identified, ~13% were predicted to bind MHCI, a small fraction of which were confirmed by mass spectrometry. The peptides were then structurally modelled bound to MHCI. Mutations that were solvent-exposed and therefore accessible to T-cell antigen receptors were predicted to be immunogenic. Vaccination of mice confirmed the approach, with each predicted immunogenic peptide yielding therapeutically active T-cell responses. The predictions also enabled the generation of peptide–MHCI dextramers that could be used to monitor the kinetics and distribution of the anti-tumour T-cell response before and after vaccination. These findings indicate that a suitable prediction algorithm may provide an approach for the pharmacodynamic monitoring of T-cell responses as well as for the development of personalized vaccines in cancer patients.

Although CD8 T cells can recognize tumour cells and mediate tumour regression following immunotherapy[10], the antigens driving effective anti-tumour CD8 T-cell responses remain largely unknown. Tumour antigens can be classified into two categories: tumour-associated self-antigens (for example, cancer-testis antigens, differentiation antigens) and antigens derived from tumour-specific mutant proteins. Since the presentation of self-antigens in the thymus may result in the elimination of high-avidity T cells, mutant neo-antigens seem likely to be more immunogenic. However, identifying these antigens has proved problematic, having evaded identification by mass spectrometry, which typically relies on sequence elucidation using public proteomic databases that do not contain patient-specific mutations. Conversely, relying on transcriptomic or exome-sequence analysis for mutation identification followed by MHCI binding prediction algorithms typically yields too many candidate mutant peptides to be easily evaluated. The strength of our approach lies in combining these two powerful and well-established analytical tools to identify tumour-associated mutated peptides that are presented on MHCI (Fig. 1a).

We performed whole-exome sequencing on MC-38 and TRAMP-C1 mouse tumour cell lines to identify tumour-specific point mutations. Coding variants were called relative to the reference mouse genome to identify 4,285 and 949 non-synonymous variants in MC-38 and TRAMP-C1, respectively (Fig. 1b). To select for high-confidence mutations and focus on the mutations likely to be expressed in the majority of the tumour cells, we subsequently selected RNA-seq-based variants that were present at a minimum of 20% allelic frequency and overlapped with the exome-based variants. This resulted in 1,290 and 67 expressed mutations in MC-38 and TRAMP-C1, respectively. Next we identified 170 predicted neo-epitopes in MC-38 and 6 predicted neo-epitopes in TRAMP-C1 using the NETMHC-3.4 algorithm[11] (Fig. 1b, Supplementary Tables 1 and 2). Despite high density exome reads, only a small number of non-synonymous variants and predicted neo-epitopes in TRAMP-C1 were identified. This low mutational frequency may at least partially explain the low immunogenicity of TRAMP-C1 observed *in vivo* (that is, paucity of tumour-infiltrating lymphocytes (TILs)), compared to MC-38 tumours (Extended Data Fig. 1a).

Next, we conducted mass spectrometric analysis for MHCI-presented peptides and searched against the transcriptome-generated FASTA database (Fig. 1a). This revealed 2,332 unique H-2K^b epitopes and 3,907 unique H-2D^b epitopes presented in MC-38, and 1,651 unique H-2K^b epitopes and 1,980 unique H-2D^b epitopes presented in TRAMP-C1 cells (Supplementary Table 3). The reduced number of epitopes identified for TRAMP-C1 compared to MC-38 probably reflects weaker MHCI surface expression on TRAMP-C1 cells[12] (Extended Data Fig. 1b). Overall we observed that peptides derived from abundant transcripts are more likely to be presented by MHCI (Fig. 1c, Supplementary Tables 3 and 4), as observed by others[13,14].

Of the 1,290 and 67 amino acid changes in MC-38 and TRAMP-C1, respectively, only 7 (7 in MC-38 and 0 in TRAMP-C1) were found to be presented on MHCI by mass spectrometry (Table 1) after manual validation and comparison with a synthetically generated version of the peptide for spectral accuracy (Extended Data Fig. 2). All but one of these neo-epitopes were predicted to bind MHCI (half-maximum inhibitory concentration (IC_{50}) < 500 nM, Table 1). Both wild-type and mutant transcripts corresponding to the peptides were expressed by MC-38 cells (Supplementary Table 5 and Extended Data Fig. 3) and although most of the exact counterpart wild-type peptides were also predicted to bind MHCI, only three were detected by mass spectrometry (Extended Data Fig. 4). The mutations were not coded in the germline as the sequence for all 7 genes was confirmed as wild type at the position of interest in the genome of a C57BL/6 mouse. These proteins have no known role in oncogenesis except Med12, which is frequently mutated in prostate cancer and in smooth muscle tumours[15,16].

The small number of mutated peptides identified by mass spectrometry compared to the number of peptides predicted to be presented may be attributed in part to the sensitivity of the peptide purification and mass spectrometric approach, but also suggests that a limiting factor to presentation could be the peptide generation and transport into the endoplasmic reticulum[17]. On the other hand, mass spectrometry allows

[1]Genentech, South San Francisco, California 94080, USA. [2]Immatics Biotechnologies GmbH, 72076 Tubingen, Germany.
*These authors contributed equally to this work.
§These authors jointly supervised this work.

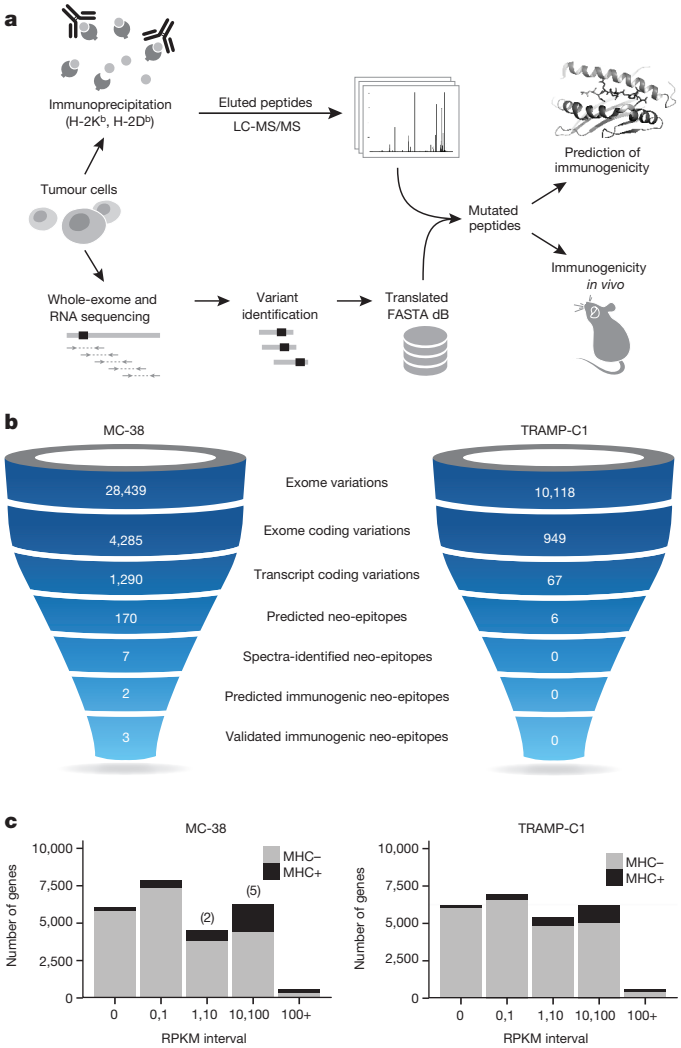


Figure 1 | Identification of MHCI-presented tumour-specific mutations in MC-38 and TRAMP-C1 tumour cell lines. **a**, Schematic of the approach for identifying mutated MHCI-presented peptides using sequencing in combination with mass spectrometry. After immunoprecipitation peptides were analysed using tandem mass spectrometry and were searched against a customized FASTA database based on RNA-seq. Immunogenicity of mutated peptides was further predicted *in silico* and validated using *in vivo* immunogenicity studies. **b**, Flowcharts for the number of genomic variations or variant peptides identified at each stage of analysis, and finally the number of peptides validated as immunogenic. **c**, RNA-seq expression profiles of genes corresponding to MHCI-presented peptides. The *x* axis represents expression levels (categorized into five reads per kilobase per million reads mapped (RPKM) intervals). The numbers of genes for each RPKM category that manifest MHCI-presented peptide(s) are shown (MHC+), as well as the rest of the genes (MHC−). The numbers of mutated MHCI-presented peptides identified by mass spectrometry in each RPKM interval are indicated above the bars.

a stringent filter to be incorporated into the workflow by selecting for peptides with sufficient expression and presentation by MHCI, therefore ensuring only the peptides most likely to yield an immunological response are further investigated.

We next asked if immunogenicity of the identified neo-epitopes could be predicted. Although there is a correlation between peptide binding affinity for MHCI and immunogenicity, other factors also contribute[18]. For example, interaction of the mutated amino acid with the T-cell antigen receptor (TCR) is likely to be essential for the recognition of the mutated peptide as 'foreign'. This is especially true when the counterpart wild-type peptide is also presented on MHCI. Five out of the seven neo-epitopes exhibited high binding affinity prediction (Table 1, $IC_{50} < 50$ nM). The other neo-epitopes exhibited lower binding affinity prediction, suggesting that they might be less immunogenic. We next used published crystal structures of H-2D^b and H-2K^b and a Rosetta-based algorithm[19] to model each of the mutant peptides in complex with MHCI and analyse the potential for the mutant residue in each neo-epitope to interact with the TCR. In general, TCR recognition of displayed peptides is mediated by interactions with peptide residues 3 through 7 (ref. 20). Among the peptides with high binding scores only in the Reps1 and Adpgk peptides did the mutation lie within this range. Structure modelling also predicted that the mutated residues were oriented towards the solvent interface, and were thus judged to have good potential to be immunogenic (Table 1 and Fig. 2)[20]. On the other hand, the mutations in the Irgq, Aatf and Dpagt1 neo-epitopes were found near the carboxy-terminal end of the peptide, which probably falls outside of the TCR binding region and suggests that these neo-epitopes were unlikely to be immunogenic despite the increase in predicted binding affinity to MHCI in comparison to the counterpart wild-type peptide (Table 1 and Fig. 2).

We next evaluated the immunogenicity of mutated tumour antigens *in vivo* by immunizing C57BL/6 mice with peptides encoding the mutated epitopes in combination with adjuvant and measured CD8 T-cell responses using peptide–MHCI dextramers. Three out of six peptides elicited CD8 T-cell responses (Fig. 3a). We had predicted Reps1 and Adpgk to be immunogenic on the basis of structure and binding affinity prediction, and both elicited strong CD8 T-cell responses. Of the four peptides predicted to be non-immunogenic, only Dpagt1 induced a weak CD8 T-cell response. The exact counterpart wild-type peptides of Reps1 and Adpgk were not immunogenic (Extended Data Fig. 5a–d).

We confirmed the immunogenicity of these mutated peptides in the context of the MC-38 tumours grown in C57BL/6 mice by analysing TILs and staining with peptide–MHCI dextramers. We found that CD8 T cells specific for Reps1, Adpgk and Dpagt1 were enriched in the tumour bed (Fig. 3b) but not T cells specific for the other mutant peptides (data not shown). Although there was heterogeneity, Adpgk-specific CD8 T cells were the most abundant of the three, and this was specific to MC-38 tumours as no Adpgk-specific CD8 T cells were detected in TRAMP-C1 tumours (Extended Data Fig. 5e). These results show that CD8 T-cell responses are generated against neo-epitopes in MC-38 tumours and further supports our hypothesis that structural correlation predicting solvent exposure of the mutation combined with MHCI-binding affinity may provide suitable criteria for triaging potential

Table 1 | Summary of mutant peptides presented by MHCI in the MC-38 cell line

Gene	Peptide	MHC allele	IC_{50} (mut)	IC_{50} (WT)	Mutation position	Immunogenicity prediction
Dpagt1	SIIVFNL[V/**L**]	*H-2K*^{b}	8	34	Anchor (P8)	–
Reps1	AQL[P/**A**]NDVVL	*H-2D*^{b}	9	100	Solvent (P4)	+
Adpgk	ASMTN[R/**M**]ELM	*H-2D*^{b}	2	3	Solvent (P6)	+
Cpne1	SSP[D/**Y**]SLHYL	*H-2D*^{b}	211	685	Solvent (P4)	–
Irgq	AALLNSA[G/**V**]L	*H-2D*^{b}	3	52	Solvent (P8)	–
Aatf	MAPIDHT[A/**T**]M	*H-2D*^{b}	30	102	Solvent (P8)	–
Med12	DPSSSVLFE[D/**Y**]	*H-2K*^{b}	38,300	39,411	No structure	–

The IC_{50} values (IC_{50} predicted by NetMHC-3.4 in nM units) are shown. Peptides with $IC_{50} < 500$ nM are predicted to bind MHCI. The mutant amino acid's position is described (whether it was anchored to a MHC molecule or solvent-exposed), based on homology to known MHC–peptide complex structures. The immunogenicity prediction is simply based on $IC_{50} < 50$ nM for the peptide, and position (P3–7: which are likely to interact with TCR) and solvent exposure of the mutation. Mutation is indicated in bold next to the wild-type amino acid in the bracket in the peptide sequences.

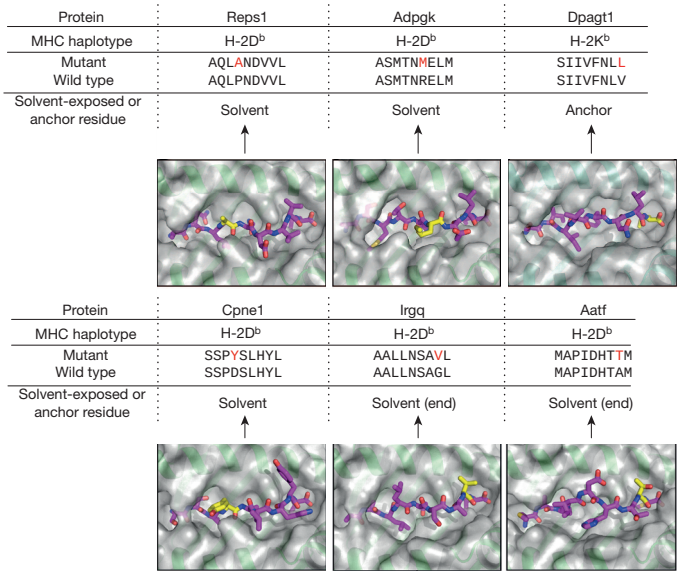


Figure 2 | Modelling of mutant peptide–MHCI complexes. Mutant peptides were modelled into peptide–MHCI structures using existing crystal structures from the Protein Data Bank as starting models and optimizing the conformation of the bound mutant peptide using the program FlexPepDock[19]. H-2D^b and H-2K^b MHC haplotypes are shown in green. Modelled peptides are shown in magenta as stick models, with the mutated residue highlighted in yellow.

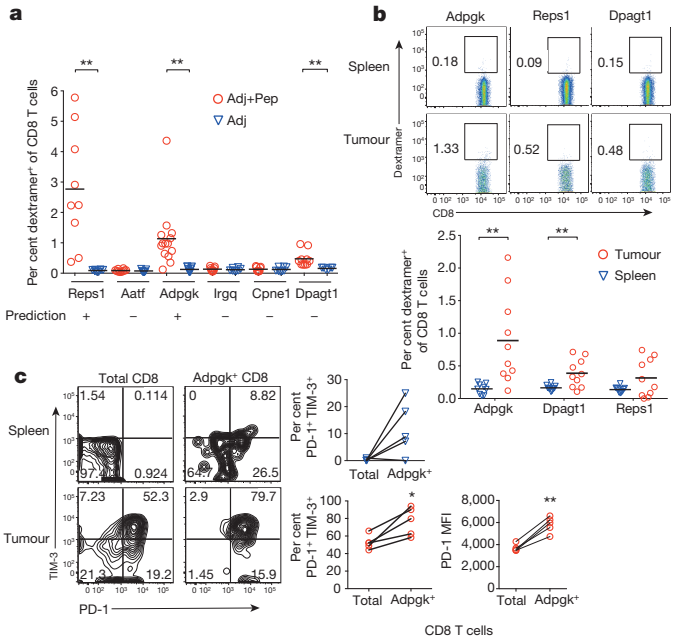


Figure 3 | Immunogenicity of mutated peptides *in vivo*. **a**, C57BL/6 mice were immunized with adjuvant (anti-CD40 antibody plus poly(I:C)) alone (Adj) or in combination with a mix of all long mutated peptides from Table 1 (50 μg each) (Adj+Pep). Med12 peptide was not included because it was not predicted to bind to MHCI. After two immunizations on day 0 and day 14, spleens were collected at day 21 and CD8 T cells were stained with different phycoerythrin (PE)-labelled peptide–MHCI dextramers (indicated on *x* axis) and analysed by flow cytometry. The adjuvant alone (Adj) group is used as the reference for immunogenicity. Each dot represents individual mouse with bar representing mean. Pooled data from two independent experiments are shown with $n \geq 6$ (Adj) or $n \geq 9$ (Adj+Pep) mice per group. **b**, Frequency of Adpgk, Reps1 or Dpagt1–MHCI-specific CD8 T cells among total CD8 T cells in MC-38 tumour-bearing mice (tumour size ~500 mm^3). The graph on the bottom represents pooled tumour data from ten individual mice. The bars represent mean, $^{**}P \leq 0.01$ (two-tailed unpaired *t*-test). **c**, PD-1 and TIM-3 surface expression was measured on Adpgk–H-2D^b dextramer$^+$ or total CD8 T cells from panel **b**. A representative staining from spleen (top) and tumour (bottom) is shown with the graph on right showing pooled data from five mice. For tumours, PD-1 median intensity of fluorescence (MFI) (gated on PD-1$^+$ cells) in the indicated population is also shown. $^*P \leq 0.05$, $^{**}P \leq 0.01$ (two-tailed paired *t*-test).

immunogenic peptides. Further analysis of a larger peptide data set will give a better indication of the statistical utility of this method.

Bulk TILs are usually analysed to monitor frequency and the phenotype of anti-tumour CD8 T-cell responses, which may not provide a true assessment because only a fraction of CD8 TILs are tumour-specific[6]. Using peptide–MHCI dextramers for the three immunogenic peptides, we were able to compare the anti-tumour TILs with the bulk TILs in MC-38 tumours. Although the frequency of tumour-specific CD8 T cells infiltrating tumours appear to increase initially, we could not discern a clear correlation between infiltration and tumour growth (Extended Data Fig. 6). Interestingly, a vast majority (76.9 ± 7.1% s.e.m.) of tumour-specific CD8 TILs, detected using Adpgk–H-2D^b dextramer, co-expressed PD-1 and TIM-3, compared to bulk CD8 TILs (52.6 ± 3.6%) (Fig. 3c). PD-1 and TIM-3 are inhibitory receptors expressed upon T cell activation and their coexpression on CD8 T cells is associated with chronic antigen stimulation leading to T cell exhaustion[21–23]. We also observed a population of CD8 TILs (Adpgk–H-2D^b dextramer-negative) that expressed PD-1 and TIM-3. This could be due to the effect of the tumour microenvironment on CD8 T cells irrespective of their specificity[24–26]. However, the expression of PD-1 on bulk CD8 TILs was lower than tumour Adpgk-specific CD8 T cells (PD-1 MFI (median fluorescence intensity), Fig. 3c). These results highlight the differential phenotypic characteristics of tumour-reactive cells in the tumour microenvironment and suggest that tumour-specific CD8 T cells exhibit a more exhausted state, in line with recent observations in human tumours[6].

To determine if CD8 T cells induced against neo-epitopes could provide protective anti-tumour immunity, healthy mice were immunized with the mutated peptides (Adpgk, Reps1 and Dpagt1) and subsequently challenged with MC-38 tumour cells (Fig. 4a). Tumour growth was completely inhibited in most of the animals in the vaccine group compared to adjuvant alone (Fig. 4b). Two of the animals that had substantial tumour growth in the vaccine group also had the lowest frequency of Adpgk-reactive CD8 T cells in blood before tumour inoculation (Fig. 4c and Extended Data Fig. 7), strongly supporting that CD8 T-cell responses specific to mutated peptides conferred protection.

Next, we asked if the neo-epitope-specific CD8 T-cell responses could be further amplified in tumour-bearing mice upon immunization (Fig. 4d). After a single immunization, the frequency of Adpgk-reactive CD8 T cells increased remarkably in the spleen of tumour-bearing mice compared to immunized non-tumour-bearing animals (Fig. 4e). Interestingly, we did not observe a similar enhancement of the CD8 T-cell response against Reps1 and Dpagt1 (Extended Data Fig. 8), suggesting that Adpgk is the immuno-dominant neo-epitope in tumour-bearing mice. We also observed a nearly threefold increase in accumulation of Adpgk-reactive CD8 T cells among total CD8 TILs in the tumours (Fig. 4e). Peptide vaccination also increased overall infiltration of CD45$^+$ cells and CD8$^+$ T cells in tumours, which resulted into nearly 20-fold increase in the frequency of neo-epitope-specific CD8 T cells among the total live cells in the tumour (Fig. 4f and Extended Data Fig. 9a).

We next analysed the phenotype of peptide-specific CD8 T cells induced by peptide vaccination. We found that the frequency of PD-1 and TIM-3 expressing Adpgk-specific CD8 TILs was reduced after vaccination, and the surface expression of PD-1 and TIM-3 on these cells was also reduced (Fig. 4g, h). This might be an adjuvant effect as it was also seen in the adjuvant alone group. These results suggest that tumour-specific T cells exhibit a less exhausted phenotype after vaccination, and this was further confirmed by the higher percentage of IFN-γ-expressing CD8 and CD4 TILs in the tumours from vaccinated mice (Fig. 4i and Extended Data Fig. 9b).

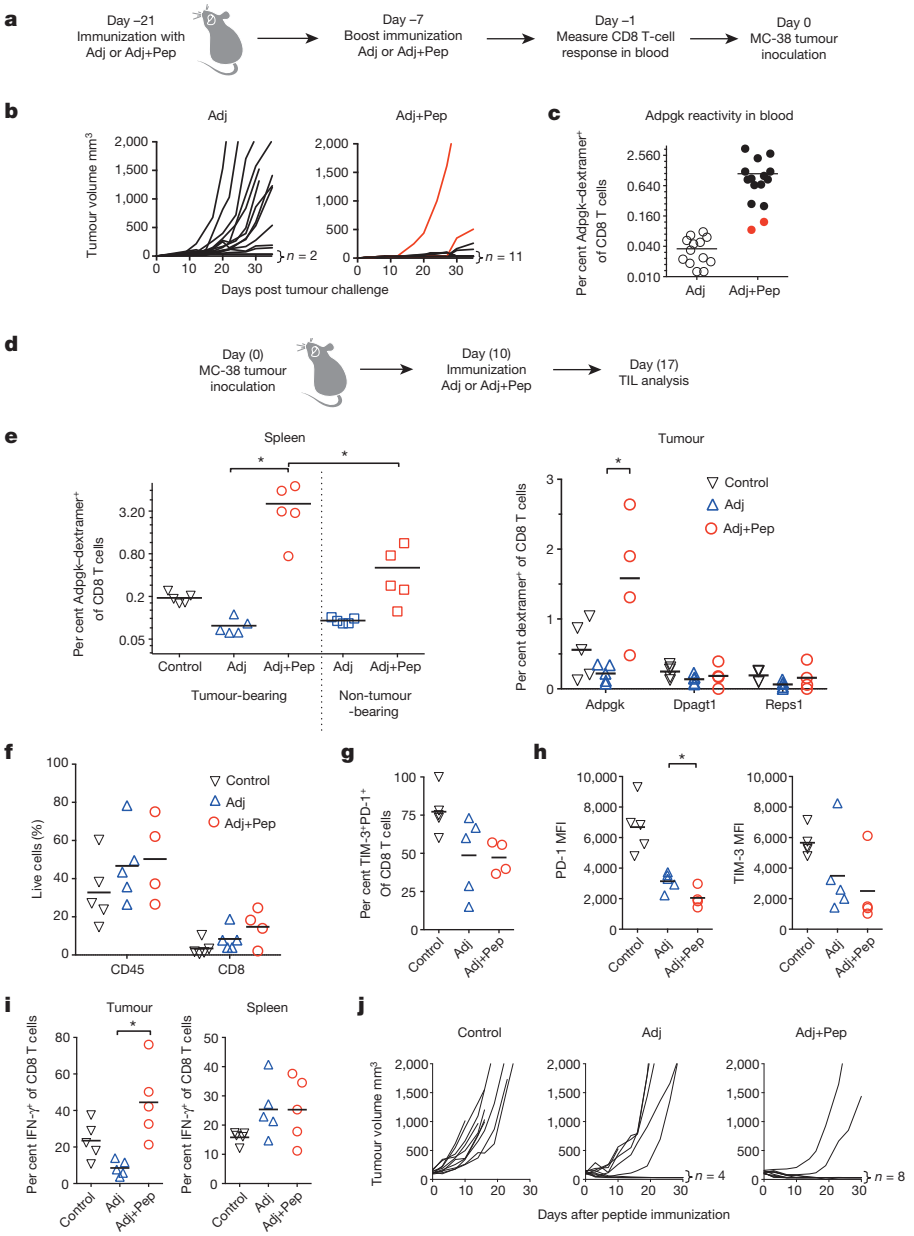


Figure 4 | Vaccination with immunogenic peptides provides protection and delays tumour growth. **a**–**c**, C57BL/6 mice were immunized with mutated peptides (Adpgk, Reps1 and Dpagt1, 50 μg each) and anti-CD40 antibody plus poly(I:C) as adjuvant (Adj+Pep) or adjuvant alone (Adj) followed by inoculation with MC-38 tumour cells as outlined in **a**. **b**, Tumour growth for Adj ($n = 14$) or Adj+Pep ($n = 15$) groups. Two mice in the Adj+Pep group that showed tumour growth are shown in red. **c**, Frequency of CD8 T cells in blood (at day −1) stained with Adpgk–H-2D^b dextramer and analysed by flow cytometry. The red dots in the graph correspond to the mice that developed tumours in the Adj+Pep group (in **b**). Pooled data from two independent experiments are shown. **d**–**j**, C57BL/6 mice were inoculated with MC-38 tumour cells, and 10 days later immunized with mutated peptides (Adpgk, Reps1 and Dpagt1, 50 μg each) with adjuvant (Adj+Pep), adjuvant alone (Adj) or left untreated (control) as outlined in **d**. **e**, Seven days post immunization, spleen or tumour cells were stained with PE-labelled peptide–MHCI dextramers for the indicated peptide (for tumours) or with Adpgk–H-2D^b dextramer (for spleen) and analysed by flow cytometry. For comparison, non-tumour-bearing mice were also immunized with adjuvant alone (Adj) or adjuvant with peptides (Adj+Pep) and splenic CD8 T cells were stained with PE-labelled Adpgk–H-2D^b dextramer at day 7. **f**, CD45$^+$ or CD8$^+$ T cells in tumours shown as percentage of total live cell gate. **g**, **h**, Frequency of TIM-3$^+$PD-1$^+$ cells and mean intensity of fluorescence (MFI) of PD-1 (on PD-1$^+$) and TIM-3 (on PD-1$^+$TIM-3$^+$) among Adpgk–H-2D^b dextramer$^+$ CD8 TILs measured by flow cytometry. **i**, TILs or splenocytes were stimulated with PMA and ionomycin and IFN-γ production in CD8 T cells was determined by intracellular cytokine staining and analysis by flow cytometry. **j**, Tumour growth in control, Adj or Adj+Pep groups ($n = 10$ each group). **e**–**j**, Data are representative of two independent experiments. The bars represent mean. *$P \leq 0.05$ (two-tailed unpaired t-test).

Finally, even in this more difficult therapeutic setting, tumour-bearing mice vaccinated with mutated peptides showed remarkable and sustainable inhibition of tumour growth compared to untreated control or adjuvant alone groups (Fig. 4j). Thus, simple peptide vaccination with the predicted neo-epitopes generated sufficient T-cell immunity to reject a previously established tumour.

The identification of epitopes that drive the immune response in cancer is essential to the understanding and manipulation of CD8 T-cell immune responses for clinical benefit. Recent studies in mice and humans have suggested that tumour-specific mutations probably have a key role in shaping the anti-tumour response; however, their identification has remained a challenge[2–9]. We developed a novel strategy to identify neo-epitopes by combining whole-exome/transcriptome sequencing and mass spectrometry analysis, along with a structural prediction algorithm to predict MHCI peptide immunogenicity. This will not only facilitate generation of novel vaccines but also make it feasible to track tumour-specific T cells, which could be invaluable for prognosis of cancer patients in this era of cancer immunotherapy.

To be fully useful in the clinic, it will probably be necessary to simplify the approach further, so as to rely entirely on computational predictions of peptide binding rather than mass spectrometry. A purely computational approach, including a MHCI-haplotype-specific structural prediction, would require only whole-exome/transcriptome sequencing of a patient's tumour, which is beginning to be routinely determined.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 27 May; accepted 28 October 2014.

1. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).
2. Matsushita, H. *et al.* Cancer exome analysis reveals a T-cell-dependent mechanism of cancer immunoediting. *Nature* **482,** 400–404 (2012).
3. DuPage, M., Mazumdar, C., Schmidt, L. M., Cheung, A. F. & Jacks, T. Expression of tumour-specific antigens underlies cancer immunoediting. *Nature* **482,** 405–409 (2012).
4. Castle, J. C. *et al.* Exploiting the mutanome for tumor vaccination. *Cancer Res.* **72,** 1081–1091 (2012).
5. van Rooij, N. *et al.* Tumor exome analysis reveals neoantigen-specific T-cell reactivity in an ipilimumab-responsive melanoma. *J. Clin. Oncol.* **31,** e439–e442 (2013).
6. Gros, A. *et al.* PD-1 identifies the patient-specific CD8+ tumor-reactive repertoire infiltrating human tumors. *J. Clin. Invest.* **124,** 2246–2259 (2014).

7. Tran, E. *et al.* Cancer immunotherapy based on mutation-specific CD4+ T cells in a patient with epithelial cancer. *Science* **344,** 641–645 (2014).
8. Wick, D. A. *et al.* Surveillance of the tumor mutanome by T cells during progression from primary to recurrent ovarian cancer. *Clin. Cancer Res.* **20,** 1125–1134 (2014).
9. Brown, S. D. *et al.* Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *Genome Res.* **24,** 743–750 (2014).
10. Chen, D. S. & Mellman, I. Oncology meets immunology: the cancer-immunity cycle. *Immunity* **39,** 1–10 (2013).
11. Lundegaard, C. *et al.* NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11. *Nucleic Acids Res.* **36,** W509–W512 (2008).
12. Gujar, S. A., Pan, D. A., Marcato, P., Garant, K. A. & Lee, P. W. Oncolytic virus-initiated protective immunity against prostate cancer. *Mol. Ther.* **19,** 797–804 (2011).
13. Fortier, M. H. *et al.* The MHC class I peptide repertoire is molded by the transcriptome. *J. Exp. Med.* **205,** 595–610 (2008).
14. Granados, D. P. *et al.* MHC I-associated peptides preferentially derive from transcripts bearing miRNA response elements. *Blood* **119,** e181–e191 (2012).
15. Barbieri, C. E. *et al.* Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nature Genet.* **44,** 685–689 (2012).
16. Bertsch, E. *et al.* MED12 and HMGA2 mutations: two independent genetic events in uterine leiomyoma and leiomyosarcoma. *Mod. Pathol.* **27,** 1144–1153 (2014).
17. Goldberg, A. L. Functions of the proteasome: from protein degradation and immune surveillance to cancer therapy. *Biochem. Soc. Trans.* **35,** 12–17 (2007).
18. Sette, A. *et al.* The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. *J. Immunol.* **153,** 5586–5592 (1994).
19. London, N., Raveh, B., Cohen, E., Fathi, G. & Schueler-Furman, O. Rosetta FlexPepDock web server–high resolution modeling of peptide-protein interactions. *Nucleic Acids Res.* **39,** W249–W253 (2011).
20. Rudolph, M. G., Stanfield, R. L. & Wilson, I. A. How TCRs bind MHCs, peptides, and coreceptors. *Annu. Rev. Immunol.* **24,** 419–466 (2006).
21. Jin, H. T. *et al.* Cooperation of Tim-3 and PD-1 in CD8 T-cell exhaustion during chronic viral infection. *Proc. Natl Acad. Sci. USA* **107,** 14733–14738 (2010).
22. Jin, H. T., Ahmed, R. & Okazaki, T. Role of PD-1 in regulating T-cell immunity. *Curr. Top. Microbiol. Immunol.* **350,** 17–37 (2011).
23. Wherry, E. J. T cell exhaustion. *Nature Immunol.* **12,** 492–499 (2011).
24. Fourcade, J. *et al.* Upregulation of Tim-3 and PD-1 expression is associated with tumor antigen-specific CD8+ T cell dysfunction in melanoma patients. *J. Exp. Med.* **207,** 2175–2186 (2010).
25. Fourcade, J. *et al.* $CD8^+$ T cells specific for tumor antigens can be rendered dysfunctional by the tumor microenvironment through upregulation of the inhibitory receptors BTLA and PD-1. *Cancer Res.* **72,** 887–896 (2012).
26. Crespo, J., Sun, H., Welling, T. H., Tian, Z. & Zou, W. T cell anergy, exhaustion, senescence, and stemness in the tumor microenvironment. *Curr. Opin. Immunol.* **25,** 214–221 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements The authors thank A. Bruce and J. Murphy for excellent assistance with artwork.

Author Contributions M.Y. was involved in planning and performing *in vivo* experiments, analysing and interpreting data, and writing the manuscript. S.J. analysed and interpreted whole-exome sequencing and RNA sequencing data, generated translated FASTA database, searched for potential neo-epitopes. Q.T.P. and T.K.C. performed mass spectrometric data analysis and peptide validation. P.L. performed the structure prediction of the MHCI–peptide complexes. J.T. performed studies with tumour-bearing mice. S.B. performed and analysed FACS studies on tumour lines. C.F. performed and analysed immunizations experiments. Z.M. oversaw RNA sequencing experiments. I.M. assisted with the study design and the preparation of the manuscript. J.F. and T.W. performed MHCI peptide isolation and mass spectrometric analysis. L.D and J.R.L. oversaw all the work performed, planned experiments, interpreted data and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to L.D. (delamarre.lelia@gene.com) or J.R.L. (lill.jennie@gene.com).

LETTER

doi:10.1038/nature13988

Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens

Matthew M. Gubin[1], Xiuli Zhang[2], Heiko Schuster[3], Etienne Caron[4], Jeffrey P. Ward[1,5], Takuro Noguchi[1], Yulia Ivanova[1], Jasreet Hundal[6], Cora D. Arthur[1], Willem-Jan Krebber[7], Gwenn E. Mulder[7], Mireille Toebes[8], Matthew D. Vesely[1], Samuel S. K. Lam[1], Alan J. Korman[9], James P. Allison[10], Gordon J. Freeman[11], Arlene H. Sharpe[12], Erika L. Pearce[1], Ton N. Schumacher[8], Ruedi Aebersold[4,13], Hans-Georg Rammensee[3], Cornelis J. M. Melief[7,14], Elaine R. Mardis[6,15], William E. Gillanders[2], Maxim N. Artyomov[1] & Robert D. Schreiber[1]

The immune system influences the fate of developing cancers by not only functioning as a tumour promoter that facilitates cellular transformation, promotes tumour growth and sculpts tumour cell immunogenicity[1–6], but also as an extrinsic tumour suppressor that either destroys developing tumours or restrains their expansion[1,2,7]. Yet, clinically apparent cancers still arise in immunocompetent individuals in part as a consequence of cancer-induced immunosuppression. In many individuals, immunosuppression is mediated by cytotoxic T-lymphocyte associated antigen-4 (CTLA-4) and programmed death-1 (PD-1), two immunomodulatory receptors expressed on T cells[8,9]. Monoclonal-antibody-based therapies targeting CTLA-4 and/or PD-1 (checkpoint blockade) have yielded significant clinical benefits—including durable responses—to patients with different malignancies[10–13]. However, little is known about the identity of the tumour antigens that function as the targets of T cells activated by checkpoint blockade immunotherapy and whether these antigens can be used to generate vaccines that are highly tumour-specific. Here we use genomics and bioinformatics approaches to identify tumour-specific mutant proteins as a major class of T-cell rejection antigens following anti-PD-1 and/or anti-CTLA-4 therapy of mice bearing progressively growing sarcomas, and we show that therapeutic synthetic long-peptide vaccines incorporating these mutant epitopes induce tumour rejection comparably to checkpoint blockade immunotherapy. Although mutant tumour-antigen-specific T cells are present in progressively growing tumours, they are reactivated following treatment with anti-PD-1 and/or anti-CTLA-4 and display some overlapping but mostly treatment-specific transcriptional profiles, rendering them capable of mediating tumour rejection. These results reveal that tumour-specific mutant antigens are not only important targets of checkpoint blockade therapy, but they can also be used to develop personalized cancer-specific vaccines and to probe the mechanistic underpinnings of different checkpoint blockade treatments.

In this study, we used two distinct progressor 3-methylcholanthrene-induced (MCA) sarcoma cell lines (d42m1-T3 and F244) and asked whether they expressed sufficient immunogenicity to be controlled by checkpoint blockade immunotherapy. Both sarcoma lines were rejected in wild-type mice treated therapeutically with anti-PD-1 and/or anti-CTLA-4 (Fig. 1a). Rejection was immunologic because (1) it was ablated by administration of monoclonal antibodies (mAbs) that either deplete $CD4^+$ or $CD8^+$ cells or neutralize interferon-γ (IFN-γ); (2) it did not occur in $Rag2^{-/-}$ mice lacking T, B and natural killer T (NKT) cells or $Batf3^{-/-}$ mice lacking $CD8\alpha^+CD103^+$ dendritic cells required for tumour antigen cross-presentation to $CD8^+$ T cells (Extended Data Fig. 1a); and (3) it induced a memory response that protected mice against rechallenge with the same tumour cells that had been injected into naive mice (Extended Data Fig. 1b, c).

On the basis of our previous success using genomics approaches to identify tumour-specific mutant antigens (TSMA) responsible for the spontaneous rejection of highly immunogenic, unedited MCA sarcomas[14], we asked whether a similar approach could identify antigens responsible for anti-PD-1-mediated rejection of d42m1-T3 progressor tumours. To increase the robustness and accuracy of our epitope predictions, we modified our method as follows: (1) mutation calls from complementary DNA capture sequencing[14] were translated to corresponding protein sequences, pipelined through three major histocompatibility complex (MHC) class I epitope-binding algorithms and a median binding affinity was calculated for each predicted epitope; (2) epitopes were prioritized on the basis of predicted median binding affinities; and (3) filters were applied to the prioritized epitope list to eliminate those predicted to be poorly processed by the immunoproteasome and to deprioritize those from hypothetical proteins or those that displayed lower binding affinity to class I than their corresponding wild-type sequences. Using this approach, many epitopes were predicted for H-$2D^b$ (49,677 9-mer and 10-mer epitopes) (Extended Data Fig. 2a) and H-$2K^b$ (44,215 8-mer and 9-mer epitopes) (Fig. 1b) based on the 2,796 non-synonymous mutations expressed in d42m1-T3[14]. Focusing on epitopes with the highest predicted binding affinity to H-$2D^b$ or H-$2K^b$, we narrowed the list down to four H-$2D^b$-binding epitopes (Extended Data Fig. 2b) and 62 H-$2K^b$-binding epitopes (Fig. 1c). Applying the aforementioned filters eliminated two predicted strong-binding H-$2D^b$ epitopes (Extended Data Fig. 2c) and 20 predicted strong-binding H-$2K^b$ epitopes (Fig. 1d) (epitope binding affinity distributions to different class I alleles are distict[15]). Based on the resulting *in-silico*-generated epitope landscape, two predominant H-$2K^b$ restricted mutant epitopes were identified by their predicted binding affinities: an A506T mutation (ITYAWTRL→ITYTWTRL) in asparagine-linked glycosylation 8 (alpha-1,3-glucosyltransferase) (Alg8) and a G1254V mutation (GGFNFRTL→VGFNFRTL) in laminin alpha subunit 4 (Lama4). Based on H-$2K^b$ consensus binding, the mutations that produce these epitopes occur at positions p4 (Alg8) and p1 (Lama4). Neither functions as an anchor residue for H-$2K^b$ (these occur

[1]Department of Pathology and Immunology, Washington University School of Medicine, 660 South Euclid Avenue, St Louis, Missouri 63110, USA. [2]Department of Surgery, Washington University School of Medicine, 660 South Euclid Avenue, St Louis, Missouri 63110, USA. [3]Department of Immunology, Institute of Cell Biology, and German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ) Partner Site Tübingen, Auf der Morgenstelle 15, 72076 Tübingen, Germany. [4]Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, 8093 Zurich, Switzerland. [5]Department of Medicine, Division of Oncology, Washington University School of Medicine, 660 South Euclid Avenue, St Louis, Missouri 63110, USA. [6]The Genome Institute, Washington University School of Medicine, 4444 Forest Park Avenue, St Louis, Missouri 63108, USA. [7]ISA Therapeutics B.V., 2333 CH Leiden, The Netherlands. [8]Division of Immunology, The Netherlands Cancer Institute, 1066 CX Amsterdam, The Netherlands. [9]Bristol-Myers Squibb, 700 Bay Road, Redwood City, California 94063, USA. [10]Department of Immunology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA. [11]Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts 02115, USA. [12]Department of Microbiology and Immunobiology, Harvard Medical School, Boston, Massachusetts 02115, USA. [13]Faculty of Science, University of Zurich, Zurich, 8093 Zurich, Switzerland. [14]Department of Immunohematology and Blood Transfusion, Leiden University Medical Center, 2333ZA Leiden, The Netherlands. [15]Department of Genetics, Washington University School of Medicine, 660 South Euclid Avenue, St Louis, Missouri 63110, USA.

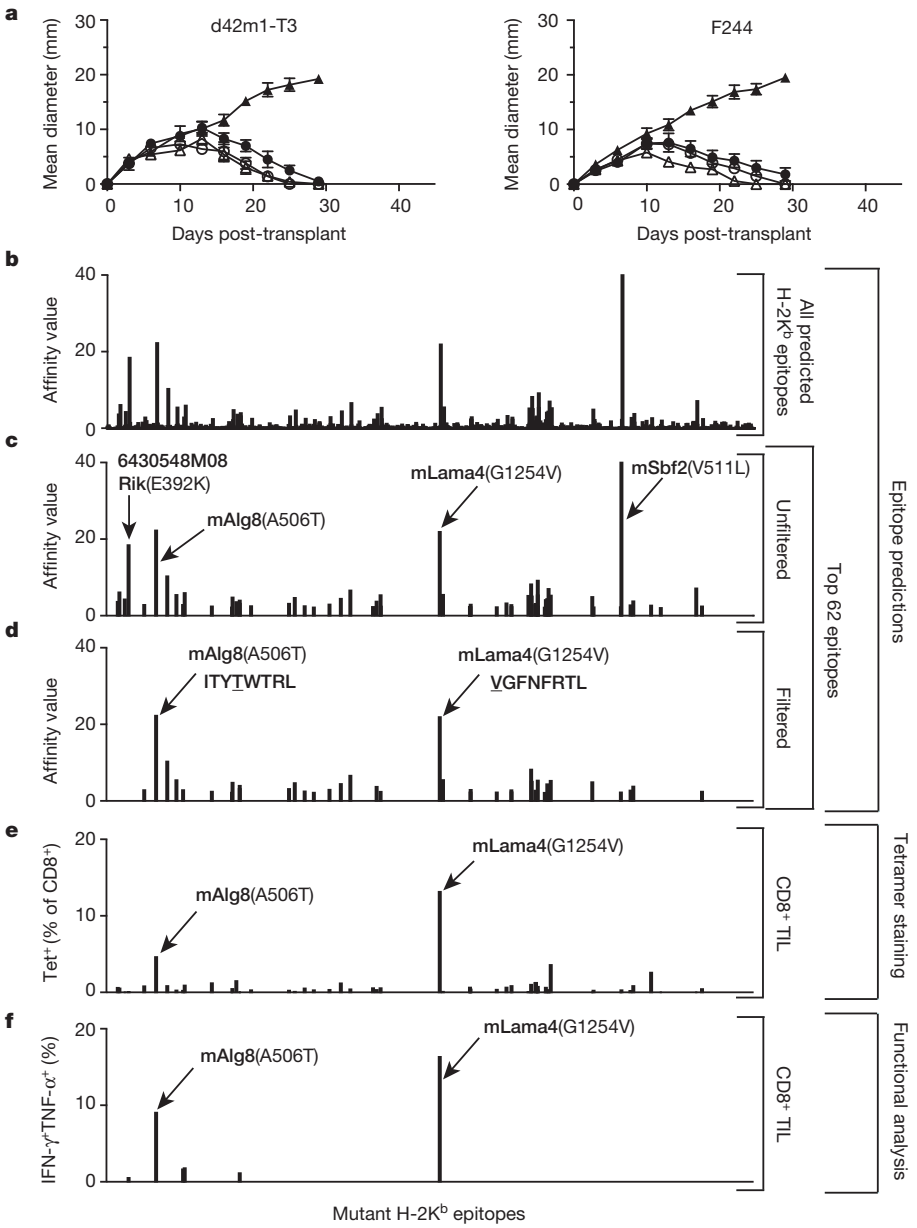


Figure 1 | Mutations in Lama4 and Alg8 form top predicted d42m1-T3 epitopes. **a**, Growth of d42m1-T3 or F244 tumours in five-mouse cohorts treated with anti-PD-1 (closed circles), anti-CTLA-4 (open circles), anti-PD-1+anti-CTLA-4 (open triangle) or control mAb (closed triangle). **b**, Potential H-2K^b binding epitopes predicted by *in silico* analysis of all missense mutations in d42m1-T3. **c**, Median affinity values for the top 62 predicted H-2K^b epitopes. **d**, Median affinity values of H-2K^b epitopes after filtering. **e**, Screening for specificities of CD8$^+$ TIL from anti-PD-1-treated, d42m1-T3-tumour-bearing mice using H-2K^b tetramers loaded with top 62 H-2K^b epitopes. **f**, IFN-γ and TNF-α induction in CD8$^+$ TIL from anti-PD-1-treated, d42m1-T3-tumour-bearing mice following culture with irradiated splenocytes pulsed with the top 62 H-2K^b peptides. Data are presented as per cent CD8$^+$ TIL expressing IFN-γ, TNF-α or for both. Data are representative of two independent experiments.

at p5 and p8). The mutant Alg8 (mAlg8) and mutant Lama4 (mLama4) epitopes are predicted to bind 1.2- and 12.8-fold stronger to H-2K^b, respectively, compared to wild-type sequences.

To identify which of the predicted d42m1-T3 neoepitopes functioned as targets for CD8$^+$ T cells in anti-PD-1-treated, tumour-bearing mice, freshly explanted CD8$^+$ tumour-infiltrating lymphocytes (TIL) were isolated just before tumour rejection (day 11) and stained with fluorescently labelled H-2K^b or H-2D^b tetramers loaded with their corresponding strong-binding 66 predicted mutant epitopes. The only tetramer-positive T cells consistently identified in the CD8$^+$ TIL population were those reacting with mLama4-H-2K^b tetramers (13.1% of CD8$^+$ TIL in experiment shown; 15.6 ± 2.7% as the mean of 6 experiments) or mAlg8-H-2K^b tetramers (4.2% of CD8$^+$ TIL in experiment shown; 2.8 ± 1.1% as the mean of 6 experiments) (Fig. 1e and Extended Data Fig. 2d). Similar results were obtained when freshly explanted CD8$^+$ TIL from the same mice were co-cultured with naive irradiated splenocytes pulsed with each of the 66 predicted H-2K^b and H-2D^b epitopes. The mLama4 and mAlg8 epitopes were, again, the only significant hits, inducing IFN-γ and tumour necrosis factor-α (TNF-α) production (Fig. 1f and Extended Data Fig. 2e). These results demonstrate that d42m1-T3 expresses two dominant TSMA epitopes for CD8$^+$ T cells following anti-PD-1 immunotherapy.

To independently validate these observations, we established CD8$^+$ T cell lines from spleens of mice that had rejected d42m1-T3 tumours after anti-PD-1 treatment. These T cells produced IFN-γ when co-cultured with d42m1-T3 but not when co-cultured with F244 or other independent sarcoma lines (Extended Data Fig. 3a). Stimulation was restricted by H-2K^b but not by H-2D^b. The only predicted epitopes that stimulated these T cell lines were mLama4 and mAlg8 (Extended Data Fig. 3b), but not their wild-type forms (Extended Data Fig. 4a).

Four subsequent findings supported the conclusion that mLama4 and mAlg8 were the relevant antigens responsible for anti-PD-1-induced rejection of d42m1-T3. First, mLama4 or mAlg8 epitopes stabilized H-2K^b expression on RMA-S cells, which lack a functional antigen transporter and thus fail to stably express MHC class I proteins on the cell surface (Extended Data Fig. 4b). Second, both epitopes were detected by mass spectrometry in eluates of affinity-purified H-2K^b isolated from d42m1-T3 tumours. Using a discovery mass spectrometry approach, we identified mLama4 in the H-2K^b eluate (Extended Data Fig. 5a) and verified its identity using an isotope-labelled synthetic mLama4 peptide

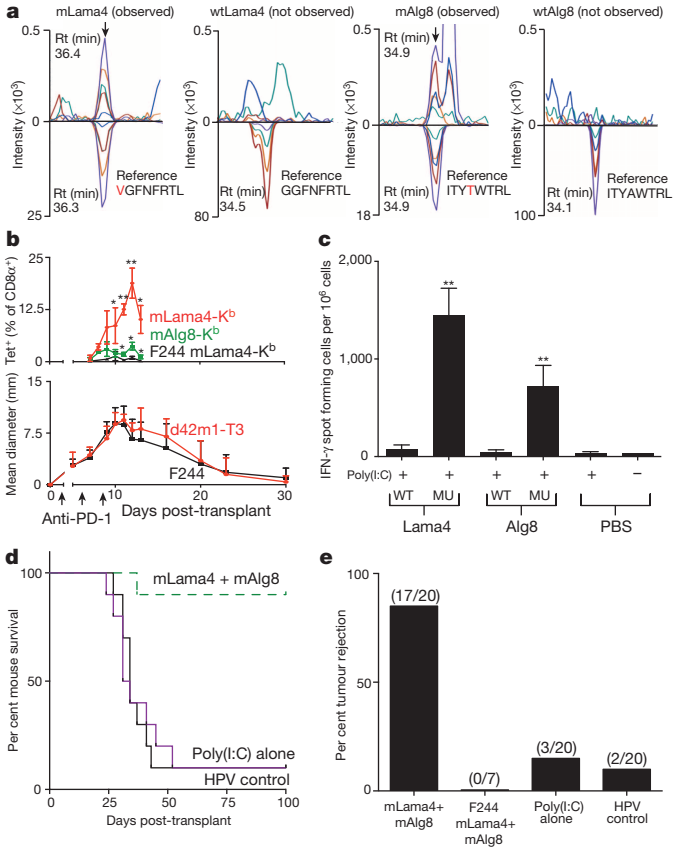


Figure 2 | Mutant Lama4 and mAlg8 are therapeutically relevant d42m1-T3 TSMA. **a**, Detection of mLama4 and mAlg8 bound to cellular H-2K^b by mass spectrometry. Rt, retention time. **b**, Top, time-dependent tumour infiltration of mLama4- and mAlg8-specific CD8$^+$ T cells ($n = 5$). Data represent means ± s.e.m. of 5 independent experiments. Bottom, growth kinetics of d42m1-T3 and F244 during anti-PD-1 immunotherapy ($n = 5$). Data represent average tumour diameter ± s.e.m. and are representative of at least three independent experiments. **c**, IFN-γ ELISPOT analysis of peptide-stimulated splenocytes from mice immunized with mLama4 or mAlg8 SLP plus poly(I:C) ($n = 3$ mice per group). Data are means ± s.e.m. Representative of two independent experiments. Samples were compared using unpaired, two-tailed Student's t test (*$P < 0.05$, **$P < 0.01$). **d**, Kaplan–Meier survival curves of d42m1-T3-tumour-bearing mice (10 mice per group) therapeutically vaccinated with SLP vaccines plus poly(I:C). mLama4 plus mAlg8 compared to HPV control: $P = 0.0002$ (log-rank (Mantel–Cox) test). Representative of two independent experiments. **e**, Cumulative data from two independent SLP therapeutic vaccine experiments using mice (7–10 per group) with d42m1-T3 or F244 tumours.

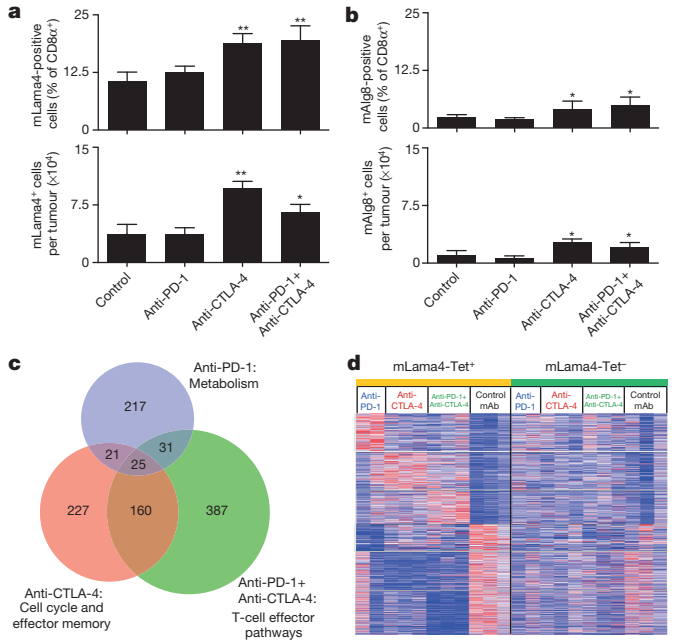


Figure 3 | Differential effects of checkpoint blockade therapy on tumour-antigen-specific CD8$^+$ T cells. **a**, **b**, Top, per cent of CD8$^+$ TIL specific for mLama4 (**a**) or mAlg8 (**b**) following checkpoint blockade therapy. Bottom, mean number of mLama4- (**a**) or mAlg8-specific (**b**) CD8$^+$ TIL per tumour following checkpoint blockade therapy. $N = 5$ mice per group pooled. Data are means ± s.e.m. of at least three independent experiments. Samples were compared to control mAb treatment using unpaired, two-tailed Student's t test (*$P < 0.05$, **$P < 0.01$). **c**, Venn diagram revealing relationships between differentially expressed genes ($P < 0.05$) in mLama4-specific CD8$^+$ TIL from mice treated with checkpoint-blocking mAbs versus control mAb. **d**, Heat map showing differentially expressed genes ($P < 0.05$) in mLama4-specific CD8$^+$ TIL from mice treated with checkpoint-blocking versus control mAbs. Colour pattern is relative with respect to the row, with red indicating gene upregulation and blue indicating gene downregulation. $n = 15$ mice per group analysed in triplicate.

(Extended Data Fig. 5b). We also found more than 200 wild-type peptides associated with H-2K^b (Supplementary Table 1), but we have no evidence that any of these function as d42m1-T3 antigens. Mutant Alg8, wild-type Lama4 and wild-type Alg8 peptides were not detected (Supplementary Table 1). In contrast, using the more sensitive targeted selected reaction monitoring (SRM) mass spectrometry method, both mLama4 and mAlg8 peptides were identified in the H-2K^b eluate (Fig. 2a, Extended Data Fig. 6a and Supplementary Data 1). Notably, mLama4 and mAlg8 were the only predicted strong-binding mutant epitopes found. Peptides from wild-type Lama4 or wild-type Alg8 were not detected. Neither mLama4 nor mAlg8 were detected in H-2K^b eluates from F244 cells (Extended Data Fig. 6b). Third, as detected by staining with H-2K^b -mLama4 or -mAlg8 tetramers, CD8$^+$ T cells specific for either antigen accumulated temporally in d42m1-T3 tumours of anti-PD-1-treated mice, reaching maximal values just before tumour rejection (Fig. 2b and Extended Data Fig. 7a). No mLama4- or mAlg8-tetramer-positive TIL were observed in F244 sarcomas from anti-PD-1-treated mice. Fourth, the two H-2K^b-restricted epitopes induced antigen-specific CD8$^+$ T cell responses in naive mice when injected together with polyinosinic-polycytidylic acid (poly(I:C)) as assessed by ELISPOT (Fig. 2c).

Since mLama4- and mAlg8-specific T cells were linked to anti-PD-1-induced d42m1-T3 rejection we asked whether a therapeutic vaccine comprised of these antigens could protect against tumour outgrowth. When 10-member groups of mice bearing established d42m1-T3 tumours were vaccinated with the combination of mLama4 (28-mer) and mAlg8 (21-mer) synthetic long peptides (SLPs) with poly(I:C), 9 rejected their tumours compared to control mice vaccinated with irrelevant human papilloma virus (HPV) (30-mer) SLP plus poly(I:C) (1/10 mice survived) or poly(I:C) alone (1/10 mice survived) (Fig. 2d and Extended Data Fig. 8a). In multiple experiments, mice vaccinated with mLama4 + mAlg8 SLP + poly(I:C) displayed an 85% survival (17/20) whereas those treated with HPV SLP + poly(I:C) or poly(I:C) alone showed 10% (2/20) and 15% (3/20) survival, respectively (Fig. 2e). Prophylactic administration of the combined mLama4 and mAlg8 SLP vaccine induced 88% survival (15/17) (Extended Data Fig. 8b, c). The combined mLama4 and mAlg8 prophylactic SLP vaccine induced superior protection compared to either SLP alone (Extended Data Fig. 8c) or compared to vaccines comprised of the minimal 8 amino acid epitopes (Extended Data Fig. 8c). The d42m1-T3-specific vaccines did not prevent outgrowth of unrelated F244 sarcomas (Fig. 2e and Extended Data Fig. 8c). These results not only demonstrate that mLama4 and mAlg8 are major antigenic targets that mediate checkpoint blockade-induced rejection of d42m1-T3 tumours, but also show that anti-PD-1 or therapeutic SLP vaccines consisting of the TSMA targeted by anti-PD-1 are similarly efficacious.

Since the aforementioned analyses were conducted using anti-PD-1-treated mice bearing d42m1-T3 tumours, we asked whether the presence of mLama4- and mAlg8-specific T cells in the TIL population was dependent on checkpoint blockade therapy. T cells specific for mLama4 or mAlg8 were detected in mice treated with control mAb or anti-PD-1 and/or anti-CTLA-4 (Fig. 3a, b and Extended Data Fig. 7b). The percentage and total number of mLama4- or mAlg8-specific $CD8^+$ TIL were similar in tumours from control mAb- and anti-PD-1-treated mice but were elevated in mice treated with either anti-CTLA-4 or anti-CTLA-4 + anti-PD-1.

The observation that mLama4- and mAlg8-specific T cells were found in tumours from mice treated with either control mAb or checkpoint blockade mAbs prompted us to assess the resultant changes in the TIL population following anti-PD-1 and/or anti-CTLA-4 treatment. We used RNA sequencing (RNA-Seq) to assess gene expression in freshly isolated, mLama4-H-$2K^b$-$tetramer^+$ TIL from groups of tumour-bearing mice treated with control mAb, anti-PD-1, anti-CTLA-4 or anti-PD-1 + anti-CTLA-4. Since mLama4-specific T cells were seven times more abundant in d42m1-T3 tumours than mAlg8-specific T cells in this series of experiments, we restricted our analysis to the former. Only a subset of 25 genes was commonly regulated (either up or down) by treatment with anti-PD-1 and/or anti-CTLA-4 (Fig. 3c and Extended Data Table 1a). This group included a subset of genes whose enhanced expression is similar to that observed in $CD8^+$ T cells from mice during acute secondary viral infection and depressed in a manner similar to that of exhausted $CD8^+$ T cells in chronic viral infection[16] (Fig. 3c, Extended Data Table 1a and Supplementary Table 2). In contrast, antigen-specific $CD8^+$ TIL isolated from anti-PD-1 and/or anti-CTLA-4 treated mice displayed mostly treatment-specific alterations of non-overlapping sets of genes involved in $CD8^+$ T cell effector functions (Supplementary Table 2). The effects of checkpoint blockade on gene expression were predominantly observed on TSMA-specific T-cells and not in other $CD8^+$ TIL (Fig. 3d).

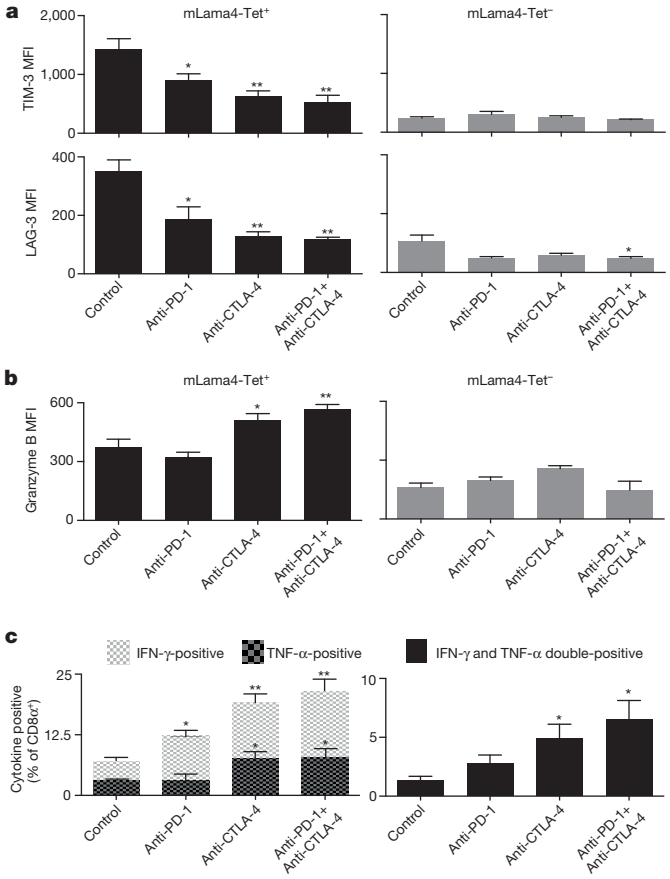


Figure 4 | Checkpoint blockade therapy alters the functional phenotypes of tumour-antigen-specific $CD8^+$ T cells. a, TIM-3 or LAG-3 expression (MFI, mean fluorescent intensity) on $CD8^+$ TIL following checkpoint blockade therapy. **b**, Granzyme B expression in $CD8^+$ TIL following checkpoint blockade therapy. **c**, Per cent of $CD8^+$ TIL positive for IFN-γ and/or TNF-α following checkpoint blockade therapy. $n = 5$ mice per group pooled. Data are means ± s.e.m. of at least three independent experiments. Samples were compared to control mAb treated mice using an unpaired, two-tailed Student's *t* test (*$P < 0.05$, **$P < 0.01$).

To determine which pathways were regulated by the different checkpoint blockade therapies, we performed gene set enrichment analysis (GSEA) using canonical-pathway- and immunological-signature-databases. When compared to mLama4-specific TIL from control mAb-treated mice, tumour-antigen-specific TIL from mice treated with anti-PD-1 and/or anti-CTLA-4 displayed a common set of alterations involving effector function, MAPK, chemokine and cytokine receptor signalling (Extended Data Table 1b and Supplementary Table 3). In contrast, mLama4-specific TIL from mice treated with anti-PD-1, anti-CTLA-4 or both mAbs displayed profound treatment-specific pathway alterations (Extended Data Table 1b and Supplementary Table 3). Treatment with anti-PD-1 produced metabolic changes including those involving oxidative phosphorylation, glycolysis, respiratory electron transport, tricarboxylic acid cycle and pentose phosphate pathways, as well as in pathways involved in IL-2 signalling. These cells also displayed a profile consistent with response to type I IFN. Treatment with anti-CTLA-4 increased NFAT and JAK-STAT signalling pathway activity, cellular proliferation/cell cycle, and activation of effector T cells. Treatment with both anti-CTLA-4 and anti-PD-1 induced a synergistic pattern of metabolic and effector T-cell-specific functions, including those involving T-cell-mediated anti-tumour activity. This was reflected in the most significant enhancement of effector molecules such as IFN-γ, Granzyme A, Granzyme B, and Fas ligand (Supplementary Table 2). Thus, whereas blockade of different inhibitory co-stimulators leads to a common biological outcome—tumour eradication—the precise mechanisms by which this outcome is achieved differ.

We also assessed changes in expression of functionally relevant proteins on/in $CD8^+$ TIL in mice undergoing treatment with different checkpoint-blocking mAbs. TIL specific for mLama4 or mAlg8 from mice treated with anti-PD-1 and/or anti-CTLA-4 displayed lower cell surface expression of lymphocyte-activation gene 3 (LAG-3) and T cell immunoglobulin and mucin protein 3 (TIM-3) than those in progressively growing tumours in control mAb-treated mice (Fig. 4a and Extended Data Fig. 9a, b). Elevated LAG-3 and TIM-3 expression is known to mark antigen-experienced, dysfunctional (that is, exhausted) $CD8^+$ T cells[16,17] in chronic viral infection. Conversely, TIL specific for mLama4 or mAlg8 from anti-CTLA-4- or anti-CTLA-4 + anti-PD-1-treated mice displayed significantly higher levels of Granzyme B than antigen-specific TIL from mice treated with either anti-PD-1 alone or control mAb (Fig. 4b). Consistent with the RNA-Seq analysis, these changes were observed predominantly in antigen-specific TIL. In addition, whereas a low percentage of $CD8^+$ TIL from mice treated with control mAb produced IFN-γ and TNF-α (Fig. 4c and Extended Data Fig. 9c), the percentage of IFN-γ-producing TIL increased following treatment of the mice with anti-PD-1 and particularly with anti-CTLA-4 or the combination of anti-CTLA-4 + anti-PD-1. TIL expressing both cytokines, which are likely to represent the most potent anti-tumour effectors, were most highly represented following treatment of tumour-bearing mice with the combination of anti-CTLA-4 + anti-PD-1.

This report documents that TSMA are targets of checkpoint blockade immunotherapy and can be used in vaccines that therapeutically induce tumour rejection as effectively as checkpoint blockade therapy. The ability to rapidly and accurately identify TSMA using genomics and bioinformatics approaches[18–20] and use them to generate MHC tetramers to identify tumour-specific T cells provides a significant advantage to the fields of tumour immunology and cancer immunotherapy. This approach has not only facilitated the identification of the antigenic targets of T cells affected by checkpoint blockade therapy[21] but

also has provided insights into the molecular changes that occur within the tumour-antigen-specific T-cell population that give rise to the anti-tumour effects of anti-PD-1 and/or anti-CTLA-4. Our findings provide some of the first experimental support for the clinical observations that (1) checkpoint blockade therapy amplifies, in some cases, pre-existing anti-tumour T-cell responses[8,9,21,22]; (2) whereas anti-CTLA-4 treatment eliminates regulatory T cells, promotes T-cell priming, and renders the host more susceptible to autoimmunity[13,22,23], anti-PD-1 promotes T-cell activation acting as a rheostat of immune effector function[9,22,24,25]; and (3) dual blockade of CTLA-4 and PD-1 is particularly effective in promoting enhanced anti-tumour effector functions[10,22,26].

The mutational loads of the MCA sarcomas used in this study are high and similar to those of ultraviolet- and carcinogen-induced human cancers. For these types of tumours, it is likely that TSMA vaccines targeting multiple antigens will be possible, thereby providing better coverage of the tumour cell population in part due to dealing more effectively with tumour heterogeneity[27,28]. Additionally, the combination of a TSMA vaccine and checkpoint blockade may facilitate the immune system's ability to recognize less immunogenic TSMA as well as shared, tumour-associated antigens (TAA) via mechanisms that mimic epitope-spreading[29]. However, recent studies have shown that human tumours containing far fewer mutations (for example, 26 mutations) can be sensitive to TSMA-based immunotherapy even for tumour-specific antigens that are targets for class II restricted $CD4^+$ T cells[30]. Our study thus provides a strong argument to actively pursue the use of TSMA as targets for cancer immunotherapy, as a means to identify patients who would best benefit from such therapy, and as components of MHC tetramers that can be used to identify tumour-specific T cells as biomarkers of successful anti-tumour responses.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 14 May; accepted 22 October 2014.

1. Shankaran, V. *et al.* IFNγ and lymphocytes prevent primary tumour development and shape tumour immunogenicity. *Nature* **410,** 1107–1111 (2001).
2. Dunn, G. P., Bruce, A. T., Ikeda, H., Old, L. J. & Schreiber, R. D. Cancer immunoediting: from immunosurveillance to tumor escape. *Nature Immunol.* **3,** 991–998 (2002).
3. Mantovani, A., Allavena, P., Sica, A. & Balkwill, F. Cancer-related inflammation. *Nature* **454,** 436–444 (2008).
4. Grivennikov, S. I., Greten, F. R. & Karin, M. Immunity, inflammation, and cancer. *Cell* **140,** 883–899 (2010).
5. Trinchieri, G. Cancer and inflammation: an old intuition with rapidly evolving new concepts. *Annu. Rev. Immunol.* **30,** 677–706 (2012).
6. Coussens, L. M., Zitvogel, L. & Palucka, A. K. Neutralizing tumor-promoting chronic inflammation: a magic bullet? *Science* **339,** 286–291 (2013).
7. Koebel, C. M. *et al.* Adaptive immunity maintains occult cancer in an equilibrium state. *Nature* **450,** 903–907 (2007).
8. Quezada, S. A., Peggs, K. S., Simpson, T. R. & Allison, J. P. Shifting the equilibrium in cancer immunoediting: from tumor tolerance to eradication. *Immunol. Rev.* **241,** 104–118 (2011).
9. Pardoll, D. M. The blockade of immune checkpoints in cancer immunotherapy. *Nature Rev. Cancer* **12,** 252–264 (2012).
10. Wolchok, J. D. *et al.* Nivolumab plus ipilimumab in advanced melanoma. *N. Engl. J. Med.* **369,** 122–133 (2013).
11. Hamid, O. *et al.* Safety and tumor responses with lambrolizumab (anti-PD-1) in melanoma. *N. Engl. J. Med.* **369,** 134–144 (2013).
12. Topalian, S. L. *et al.* Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *N. Engl. J. Med.* **366,** 2443–2454 (2012).
13. Hodi, F. S. *et al.* Improved survival with ipilimumab in patients with metastatic melanoma. *N. Engl. J. Med.* **363,** 711–723 (2010).
14. Matsushita, H. *et al.* Cancer exome analysis reveals a T-cell-dependent mechanism of cancer immunoediting. *Nature* **482,** 400–404 (2012).
15. Paul, S. *et al.* HLA class I alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity. *J. Immunol.* **191,** 5831–5839 (2013).
16. West, E. E. *et al.* Tight regulation of memory $CD8^+$ T cells limits their effectiveness during sustained high viral load. *Immunity* **35,** 285–298 (2011).
17. Wherry, E. J. T cell exhaustion. *Nature Immunol.* **12,** 492–499 (2011).
18. Castle, J. C. *et al.* Exploiting the mutanome for tumor vaccination. *Cancer Res.* **72,** 1081–1091 (2012).
19. Robbins, P. F. *et al.* Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. *Nature Med.* **19,** 747–752 (2013).
20. Fritsch, E. F. *et al.* HLA-binding properties of tumor neoepitopes in humans. *Cancer Immunol. Res.* **2,** 522–529 (2014).
21. van Rooij, N. *et al.* Tumor exome analysis reveals neoantigen-specific T-cell reactivity in an ipilimumab-responsive melanoma. *J. Clin. Oncol.* **31,** e439–e442 (2013).
22. Curran, M. A., Montalvo, W., Yagita, H. & Allison, J. P. PD-1 and CTLA-4 combination blockade expands infiltrating T cells and reduces regulatory T and myeloid cells within B16 melanoma tumors. *Proc. Natl Acad. Sci. USA* **107,** 4275–4280 (2010).
23. Brunner, M. C. *et al.* CTLA-4-mediated inhibition of early events of T cell proliferation. *J. Immunol.* **162,** 5813–5820 (1999).
24. Keir, M. E., Butte, M. J., Freeman, G. J. & Sharpe, A. H. PD-1 and its ligands in tolerance and immunity. *Annu. Rev. Immunol.* **26,** 677–704 (2008).
25. Okazaki, T., Chikuma, S., Iwai, Y., Fagarasan, S. & Honjo, T. A rheostat for immune responses: the unique properties of PD-1 and their advantages for clinical application. *Nature Immunol.* **14,** 1212–1218 (2013).
26. Duraiswamy, J., Kaluza, K. M., Freeman, G. J. & Coukos, G. Dual blockade of PD-1 and CTLA-4 combined with tumor vaccine effectively restores T-cell rejection function in tumors. *Cancer Res.* **73,** 3591–3603 (2013).
27. Spiotto, M. T., Rowley, D. A. & Schreiber, H. Bystander elimination of antigen loss variants in established tumors. *Nature Med.* **10,** 294–298 (2004).
28. Wolkers, M. C., Brouwenstijn, N., Bakker, A. H., Toebes, M. & Schumacher, T. N. Antigen bias in T cell cross-priming. *Science* **304,** 1314–1317 (2004).
29. Corbière, V. *et al.* Antigen spreading contributes to MAGE vaccination-induced regression of melanoma metastases. *Cancer Res.* **71,** 1253–1262 (2011).
30. Tran, E. *et al.* Cancer immunotherapy based on mutation-specific $CD4^+$ T cells in a patient with epithelial cancer. *Science* **344,** 641–645 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We are grateful to K. Murphy for the *Batf3*$^{-/-}$ mice, T. Hansen for providing MHC class I antibodies and the H-2K^b construct, D. Fremont for the human β2m construct, and the National Institutes of Health (NIH) Tetramer Core Facility for producing MHC class I tetramers. We also thank R. Ahmed and M. Hashimoto for the multiplex staining strategy used to define functional and dysfunctional T cells. We thank A. Bensimon, O. Schubert and P. Kouvonen for instrument maintenance and for technical support with the mass spectrometry measurements and R. Vanganipuram, M. Selby and J. Valle for generating and supplying anti-PD-1 and anti-CTLA-4 in endotoxin-free sterile form. We also thank K. Sheehan, P. Allen, G. Dunn and R. Chan for constructive criticisms and comments, all members of the Schreiber laboratory for discussions, and the many members of The Genome Institute at Washington University School of Medicine. We would also like to thank W. Song for his assistance with the bioinformatics approaches, P. Kvistborg for assistance with tetramer combinatorial coding, and Christopher Nelson for advice with peptide-MHC monomer purification. This work was supported by grants to R.D.S. from the National Cancer Institute (RO1 CA043059, U01 CA141541), the Cancer Research Institute and the WWW Foundation; to R.D.S. and W.E.G. from The Siteman Cancer Center/Barnes-Jewish Hospital (Cancer Frontier Fund); to W.E.G. from Susan G. Komen for the Cure (Promise grant); to E.R.M. from the National Human Genome Research Institute; to G.J.F. from the National Institute of Health (P50 CA101942, P01 AI054456, P50 CA101942); to A.H.S. from the National Institute of Health (P50 CA101942); and to T.N.S. from the Dutch Cancer Society (Queen Wilhelmina Research Award). E.C. is supported by a Marie Curie Intra-European Fellowship within the Seventh Framework Programme of the European Community for Research. M.M.G. was supported by a postdoctoral training grant (T32 CA00954729) from the National Cancer Institute and is currently supported by a postdoctoral training grant (Irvington Postdoctoral Fellowship) from the Cancer Research Institute. Aspects of studies at Washington University were performed with assistance by the Immunomonitoring Laboratory of the Center for Human Immunology and Immunotherapy Programs and the Siteman Comprehensive Cancer Center.

Author Contributions M.M.G. and R.D.S. were involved in all aspects of this study including planning and performing experiments, analysing and interpreting data, and writing the manuscript. X.Z. performed peptide binding experiments, helped design and perform the vaccine experiments. H.S., E.C., R.A. and H.-G.R. planned and performed the mass spectrometry analyses, interpreted the data and were involved in writing the manuscript. T.N., J.P.W., C.D.A., M.D.V., S.S.K.L. and E.L.P., participated in assessing the phenotypes of the tumour-specific T-cell lines, interpreting the data and in writing the manuscript. M.T. helped generate MHC class I multimers. A.J.K., J.P.A., G.J.F. and A.H.S. provided mAbs, helped plan the checkpoint blockade therapy experiments, and contributed to writing the manuscript. T.N.S. helped generate MHC class I multimers, analysed data and was involved in writing the manuscript. W.-J.K., G.E.M. and C.J.M.M. produced and purified the synthetic long peptides, participated in the planning of the vaccine experiments, analysed data and were involved in writing the manuscript. J.H. and E.R.M. were responsible for genomic analyses and epitope prediction and participated in writing the manuscript. W.E.G. contributed to the design and analysis of peptide binding and vaccine experiments and in writing the manuscript. Y.I. and M.N.A were responsible for optimizing the epitope prediction method, performing the RNA-sequencing analyses, analysing data and writing the manuscript. R.D.S. oversaw all the work performed.

Author Information RNA-sequencing data are available at Gene Expression Omnibus (GEO) repository at http://www.ncbi.nlm.nih.gov/geo/ (accession number GSE62771). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.D.S. (schreiber@immunology.wustl.edu).

LETTER

doi:10.1038/nature13707

Histone H2A.Z subunit exchange controls consolidation of recent and remote memory

Iva B. Zovkic[1], Brynna S. Paulukaitis[1], Jeremy J. Day[1], Deepa M. Etikala[1] & J. David Sweatt[1]

Memory formation is a multi-stage process that initially requires cellular consolidation in the hippocampus, after which memories are downloaded to the cortex for maintenance, in a process termed systems consolidation[1]. Epigenetic mechanisms regulate both types of consolidation[2–7], but histone variant exchange, in which canonical histones are replaced with their variant counterparts, is an entire branch of epigenetics that has received limited attention in the brain[8–12] and has never, to our knowledge, been studied in relation to cognitive function. Here we show that histone H2A.Z, a variant of histone H2A, is actively exchanged in response to fear conditioning in the hippocampus and the cortex, where it mediates gene expression and restrains the formation of recent and remote memory. Our data provide evidence for H2A.Z involvement in cognitive function and specifically implicate H2A.Z as a negative regulator of hippocampal consolidation and systems consolidation, probably through downstream effects on gene expression. Moreover, alterations in H2A.Z binding at later stages of systems consolidation suggest that this histone has the capacity to mediate stable molecular modifications required for memory retention. Overall, our data introduce histone variant exchange as a novel mechanism contributing to the molecular basis of cognitive function and implicate H2A.Z as a potential therapeutic target for memory disorders.

As a first step in exploring the role of H2A.Z in cognitive function, we used immunohistochemistry to confirm its expression throughout the hippocampus (Extended Data Fig. 1a–c). Next, we showed that *H2afz*, a gene encoding H2A.Z, was inhibited at 30 min ($F_{3,14} = 6.38$, $P = 0.006$) and returned to baseline levels 2 h after contextual fear conditioning in mice (Extended Data Fig. 1d). In addition, H2A.Z levels ($F_{3,9} = 5.34$, $P = 0.02$) were reduced and promoter methylation was increased ($F_{3,21} = 12.34$, $P < 0.001$) 30 min after training (Extended Data Fig. 1e, f). Although promoter methylation negatively affects transcription[4,6,7], this role is complex[13] and may not be the direct cause of H2A.Z inhibition in our study.

H2A.Z positioning around the transcriptional start site (TSS) is strongly associated with transcription[11,14,15]. Using chromatin immunoprecipitation (ChIP), we investigated H2A.Z exchange at the −1 (first nucleosome upstream of the TSS) and +1 (first nucleosome downstream of the TSS) nucleosomes of memory-associated genes during consolidation (Fig. 1). At 30 min after training, H2A.Z binding was reduced at the +1 nucleosome of memory-promoting genes (*Npas4*: Welch's $F_{3,4.98} = 67.10$, $P < 0.001$; *Arc*: Welch's $F_{3,6.8} = 153.95$, $P < 0.001$; *Egr1*: Welch's $F_{3,7.86} = 282.71$, $P < 0.001$; *Egr2*: $F_{3,18} = 3.50$, $P = 0.04$; *Fos*: $F_{3,9} = 39.61$, $P < 0.0001$), and the expression of corresponding genes was increased during this time (*Npas4*: $F_{3,15} = 22.38$, $P < 0.001$; *Arc*: $F_{3,15} = 16.34$, $P < 0.001$; *Egr1*: $F_{3,15} = 12.55$, $P < 0.001$; *Egr2*: $F_{3,15} = 9.72$, $P = 0.001$; *Fos*: $F_{3,6} = 60.71$, $P < 0.001$). In contrast, H2A.Z incorporation for the memory suppressor *Ppp3ca* increased at the +1 nucleosome ($F_{3,9} = 5.83$, $P = 0.02$) when gene expression was reduced ($F_{3,17} = 4.07$, $P = 0.03$) (Fig. 1 and Extended Data Fig. 2), suggesting that H2A.Z at the +1 nucleosome restricts transcription. These findings are consistent with reports of stimulus-induced H2A.Z eviction[16–20] and evidence for the +1 nucleosome acting as a transcriptional barrier[15,21]. Given that our data are normalized to histone H3 to correct for potential changes in nucleosome occupancy, we conclude that H2A.Z eviction in particular is associated with activity-induced gene expression.

At the −1 nucleosome, H2A.Z binding increased for both memory-promoting and memory-suppressing genes (*Npas4*: $F_{3,8} = 12.89$, $P = 0.002$; *Egr1*: Welch's $F_{3,3.39} = 9.18$, $P = 0.04$; *Egr2*: $F_{3,8} = 7.47$, $P = 0.01$; *Fos*: $F_{3,8} = 8.23$, $P = 0.008$; *Ppp3ca*: $F_{3,8} = 9.20$, $P = 0.004$) at 30 min, irrespective of changes in gene expression (Fig. 1 and Extended Data Fig. 2). Various studies have associated H2A.Z binding in the −1 nucleosome with steady-state gene activity[11,12], but our data suggest that stimulus-induced changes in H2A.Z binding do not correlate with transcription at this time point.

H2A.Z binding returned to baseline levels within 2 h, except for a delayed increase in *Bdnf* exon IV expression ($F_{3,17} = 15.09$, $P < 0.001$) and a concomitant reduction in H2A.Z binding at the +1 nucleosome ($F_{3,18} = 3.72$, $P = 0.03$) (Extended Data Fig. 2). Of note, H2A.Z was evicted in context-only mice, even though gene expression increased only with context and shock pairing. This may reflect the 2 h time point, since *Bdnf IV* expression is typically elevated 1 h after training[4].

Indeed, the association between gene expression and H2A.Z binding was no longer evident 2 h after training. Whereas H2A.Z binding returned to baseline, gene expression remained elevated (*Arc*: Welch's $F_{3,9.27} = 12.16$, $P = 0.001$; *Egr1*: Welch's $F_{3,8.25} = 6.68$, $P = 0.01$; *Egr2*: Welch's $F_{3,4.77} = 11.13$, $P = 0.01$; *Fos*: $F_{3,17} = 5.54$, $P = 0.008$). For *Ppp3ca*, H2A.Z binding increased 2 h after training at −1 ($F_{3,10} = 28.35$, $P < 0.001$) and +1 ($F_{3,10} = 4.10$, $P = 0.04$) nucleosomes, even though gene expression returned to baseline. Thus, H2A.Z exchange is uncoupled from gene expression during the late stages of transcription, consistent with evidence that H2A.Z exchange is primarily involved in transcription initiation[18].

H2A.Z has been associated with both positive and negative effects on transcription[11,12,22], with acetylation having a positive effect[16,17,23]. Using acetylation as an indirect index of the transcriptional impact of H2A.Z, we found that 30 min after training, when H2A.Z exchange is most pronounced, acetylated H2A.Z (H2A.Zac) binding increased at the −1 nucleosome of *Egr1* ($F_{3,8} = 11.07$, $P = 0.03$) and *Fos* ($F_{3,8} = 3.92$, $P = 0.05$). Consistent with H2A.Z eviction from the +1 nucleosome at 30 min, a subset of genes also exhibited reduced acetylation at the +1 nucleosome at this time point (*Egr2*: $F_{3,8} = 4.03$, $P = 0.05$; *Egr1*: $F_{3,8} = 14.13$, $P = 0.001$) (Extended Data Fig. 3).

To directly investigate the involvement of H2A.Z in memory, we conducted adeno-associated virus (AAV)-mediated H2A.Z depletion in the pyramidal cell layer in the dorsal CA1 region of the hippocampus (Fig. 2a, b). The construct effectively reduced H2A.Z messenger RNA ($t_{17} = -4.76$, $P < 0.0001$) and produced a 55.8% reduction in H2A.Z protein levels (Fig. 2c). H2A.Z depletion was associated with improved fear memory, as evidenced by increased freezing 24 h ($t_{30} = 2.28$, $P = 0.04$) and 30 days ($t_9 = -2.31$, $P = 0.05$) after training, compared to mice injected with a scramble control (Fig. 2d, e).

To investigate the basis for improved memory, we quantified the effect of H2A.Z depletion on training-induced gene expression. H2A.Z depletion increased *Bdnf* exon IV (virus (scramble or *H2afz*)–training

[1]Department of Neurobiology and Evelyn F. McKnight Brain Institute, University of Alabama at Birmingham, Birmingham, Alabama 35294, USA.

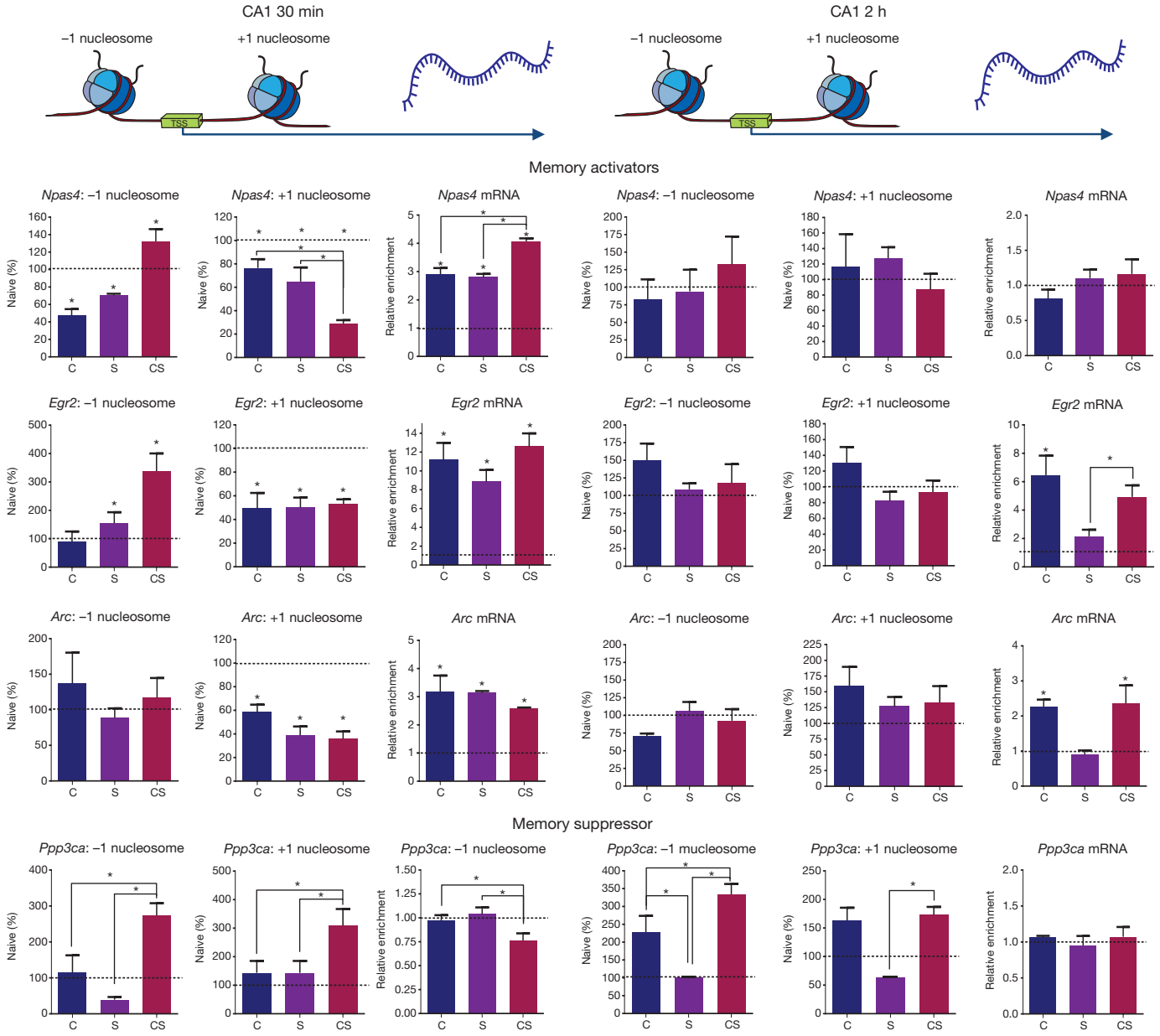


Figure 1 | H2A.Z exchange in CA1. H2A.Z binding at the −1 nucleosome (first column for each time point; *n* mice per group: N = 4, C = 3, S = 3, CS = 3) and +1 nucleosome (second column for each time point) relative to TSS either 30 min (left; *n* mice per group for *Npas4*, *Egr2* and *Arc*: N = 7; C = 5; S = 4; CS = 6; *Ppp3ca*: N = 4, C = 3, S = 3, CS = 3) or 2 h (right; *n* mice per group: N = 10; C = 2; CS = 4; S = 6; for *Ppp3ca*: N = 6; C = 2; S = 2; CS = 4) after training. Corresponding gene expression is shown in the third column for each time point (*n* mice per group: N = 5, C = 6; S = 2; CS = 6). N, naive; C, context; S, shock; CS, context plus shock. Data are expressed as mean percentage ± standard error of the mean (s.e.m.) relative to the mean of naive mice. *Follow-up comparisons with $P < 0.05$.

(naive or fear conditioned) interaction ($F_{1,17} = 8.31$, $P = 0.01$)) and *Arc* (virus–training interaction; $F_{1,17} = 6.03$, $P = 0.025$) expression 30 min after training, whereas the expression of other memory-promoting genes increased only as a function of fear conditioning, irrespective of H2A.Z manipulation (main effect of training (*Npas4*: $F_{1,12} = 7.12$, $P = 0.02$; *Egr1*: $F_{1,17} = 16.23$, $P = 0.001$; *Egr2*: $F_{1,12} = 38.53$, $P < 0.001$; *Fos*: $F_{1,12} = 93.69$, $P < 0.001$)). The memory suppressor genes *Ppp3ca* and *Ppp1cc* were not altered by training or by H2A.Z manipulation, suggesting that the effects of H2A.Z depletion are gene-specific (Fig. 2g–n). These data are consistent with evidence that H2A.Z depletion increases gene expression[19,20,22], but it is not clear which specific genes account for the memory-enhancing effects of H2A.Z depletion.

To examine the genome-wide transcriptional impact of H2A.Z depletion, we performed directional, poly(A)$^+$ RNA sequencing. H2A.Z knockdown altered baseline expression of 451 genes (Fig. 3a and Supplementary Table 1). In H2A.Z-depleted mice, fear conditioning altered the expression of 202 genes (Fig. 3b and Supplementary Table 2), including a number of the early learning-related genes identified via quantitative real-time PCR (qPCR; *Arc*, *Fos*, *Egr1* and *Egr2*; Fig. 3c, e). Thus, whereas H2A.Z knockdown altered the baseline expression of 153 genes that are affected by training, our whole-genome sequencing results are consistent with a lack of H2A.Z involvement in baseline expression of known memory-related genes (for example, *Arc*). Gene ontology analysis of training-induced changes identified genes involved in sequence-specific DNA binding and DNA regulation (Fig. 3d), consistent with the rapid transcriptional role of H2A.Z following learning.

Whereas initial memory consolidation is dependent on the hippocampus, epigenetic modifications in the cortex are implicated in systems consolidation and memory maintenance[2,6]. In contrast to the hippocampus, we did not find differences in cortical H2A.Z expression after

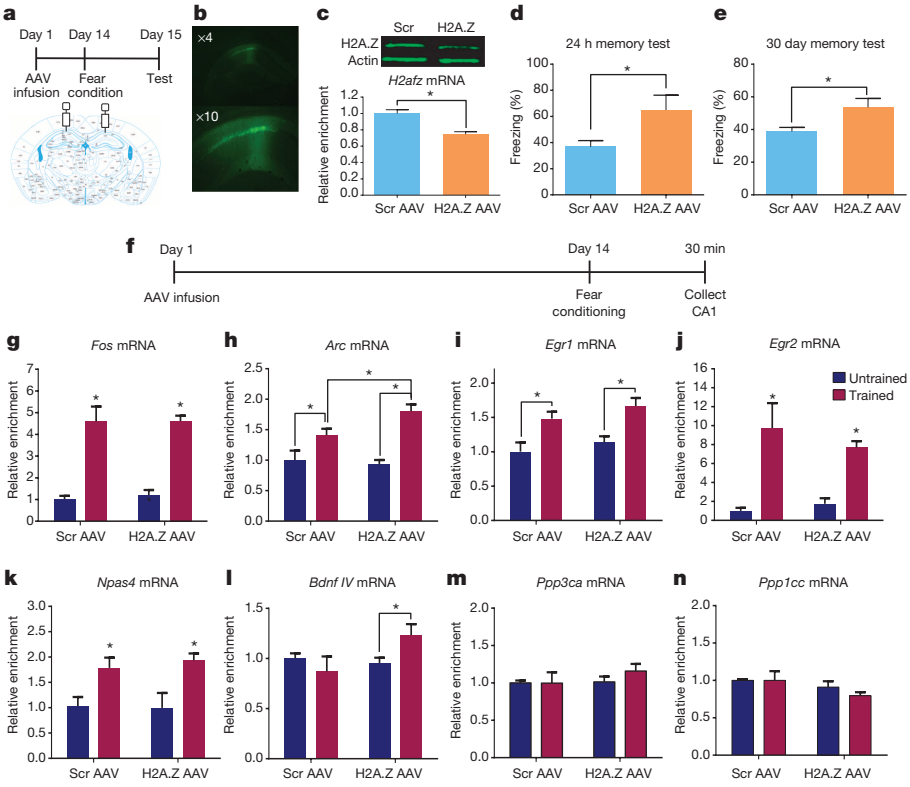


Figure 2 | H2A.Z depletion in CA1. **a**, Design of behavioural experiments. **b**, AAV spread. **c**, *H2afz*-AAV knockdown (Scrambled (scr) $n = 9$; H2A.Z $n = 10$). **d**, **e**, *H2afz*-AAV enhanced memory 24 h (**d**; Scr $n = 18$; H2A.Z $n = 14$) and 30 days (**e**; Scr $n = 5$; H2A.Z $n = 6$) after training. **f**, Design of gene expression experiments. **g**–**n**, Effect of *H2afz*-AAV on mRNA levels of memory-associated genes in untrained and trained mice (n mice per group for *Fos*, *Egr2*, *Ppp3ca*, *Ppp1cc*, *Npas4*: scr, untrained $n = 3$; scr, trained $n = 3$; H2A.Z, untrained $n = 7$; H2A.Z, trained $n = 7$; n per group for *Bdnf IV*, *Egr1* and *Arc*: scr, untrained $n = 6$; scr, trained $n = 3$; H2A.Z, untrained $n = 8$; H2A.Z, trained $n = 4$). Data expressed as mean ± s.e.m. *Follow-up comparisons with $P < 0.05$.

fear conditioning (Extended Data Fig. 4). However, H2A.Z binding at the +1 nucleosome was reduced 2 h after training (*Arc*: $F_{3,12} = 4.05$, $P = 0.03$; *Egr1*: $F_{3,12} = 3.53$, $P = 0.049$; *Egr2*: $F_{3,12} = 5.36$, $P = 0.01$), whereas H2A.Z binding increased at the +1 nucleosome of the memory suppressor *Ppp3ca* ($F_{3,12} = 4.06$, $P = 0.03$). Changes in H2A.Z binding at the −1 nucleosome were found only for *Ppp3ca* ($F_{3,12} = 13.84$, $P < 0.001$), where less H2A.Z was present at 2 h (Extended Data Figs 5 and 6). Training-induced H2A.Z eviction at 2 h implicates H2A.Z in the early stages of systems consolidation.

At 7 days, when memory becomes increasingly dependent on the cortex, H2A.Z binding increased at the −1 nucleosome of memory-promoting genes (*Arc*: $F_{3,21} = 3.03$, $P = 0.05$; *Egr1*: $F_{3,12} = 5.46$, $P = 0.01$;

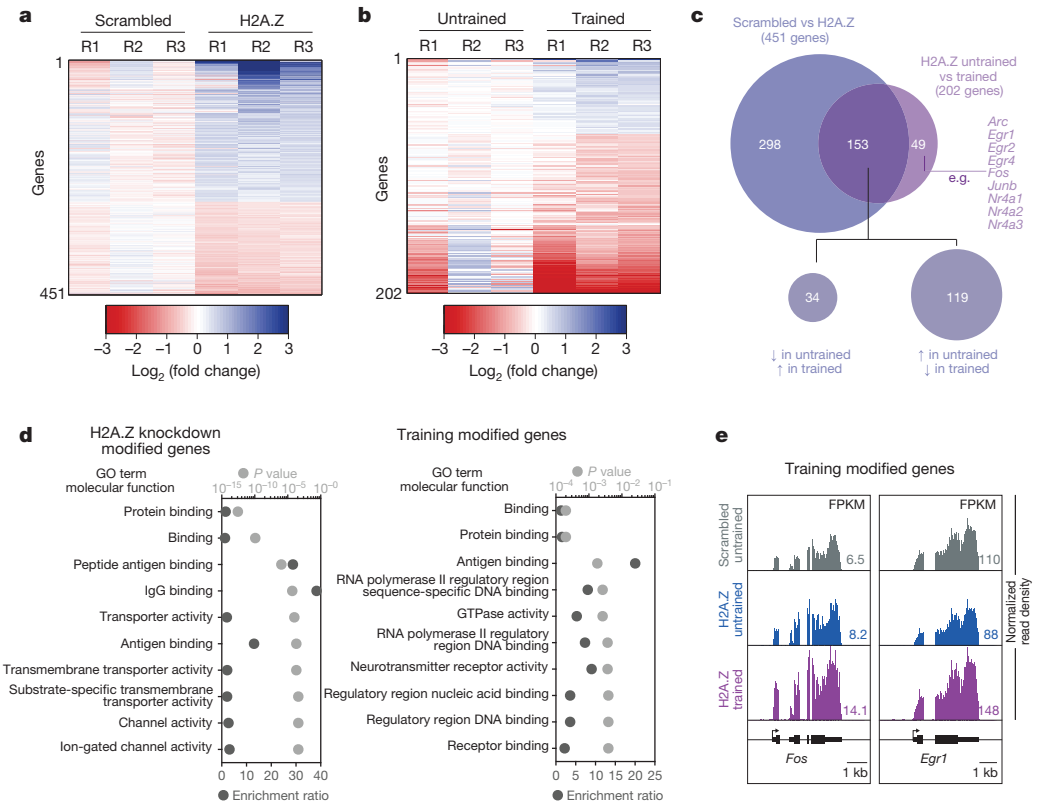


Figure 3 | RNA sequencing data depicting genome-wide transcriptional impact of AAV-mediated H2A.Z depletion. **a**, A comparison of untrained animals two weeks after stereotaxic delivery of scrambled or H2A.Z AAV into CA1. **b**, Comparison of gene expression in mice receiving H2A.Z AAV with and without fear conditioning. Samples were taken 30 min after training. **c**, Venn diagram depicting the overlap between genes mediated by H2A.Z AAV at baseline and in response to fear conditioning. **d**, Gene ontology analysis of differentially expressed genes in untrained scrambled versus H2A.Z AAV and of H2A.Z AAV mice with and without training. **e**, Example of individual genes modified by training. FPKM, fragments per kilobase of exon per million mapped reads; kb, kilobase. $n = 3$ mice per group.

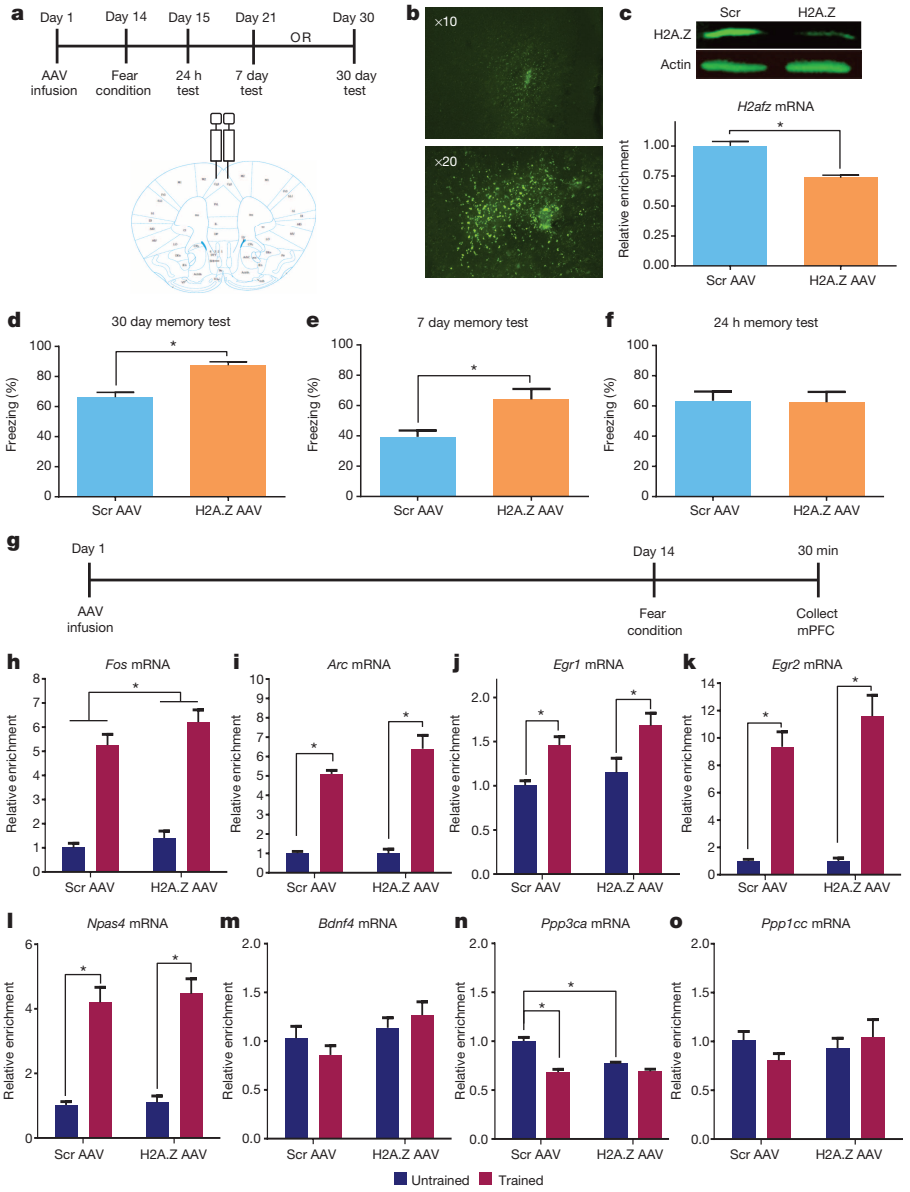


Figure 4 | H2A.Z depletion in mPFC. **a**, Experimental design for behaviour. **b**, AAV spread. **c**, H2A.Z-AAV reduced H2A.Z protein and mRNA expression ($t_3 = 6.91$, $P = 0.006$; scr $n = 2$; H2A.Z $n = 3$). **d**–**f**, Freezing behaviour at 30 days (**d**; $n = 8$ mice per group), 7 days (**e**; $n = 5$ mice per group) or 24 h (**f**; $n = 5$ mice per group) after training. **g**, Experimental design for mRNA measurement. **h**–**o**, Effect of AAV treatment on mRNA levels of memory-associated genes in untrained (naive) and fear-conditioned mice. (n mice per group: scr, untrained $n = 5$; scr, trained $n = 3$; H2A.Z, untrained $n = 4$; H2A.Z, trained $n = 4$). Data are expressed as mean ± s.e.m. *Follow up comparisons with $P < 0.05$.

Egr2: $F_{3,12} = 3.66$, $P = 0.04$; *BdnfIV*: $F_{3,12} = 4.21$, $P = 0.03$) and the −1 nucleosome of the memory suppressor *Ppp3ca* ($F_{3,21} = 5.98$, $P = 0.004$). These changes were no longer evident at 30 days (Extended Data Figs 5 and 6), indicating that TSS-flanking H2A.Z is associated with systems consolidation, but perhaps not with memory maintenance, consistent with a recent study of cortical histone acetylation[2].

In contrast to observations of cortical H2A.Z exchange at 2 h, we did not find differences in H2A.Zac binding at this time (Extended Data Fig. 7). At 7 days, H2A.Zac binding was reduced at a subset of −1 (*Egr1*: $F_{3,11} = 4.96$, $P = 0.02$; *Egr2*: $F_{3,11} = 3.58$, $P = 0.05$; *BdnfIV*: $F_{3,11} = 6.83$, $P = 0.007$; *Ppp3ca*: $F_{3,11} = 4.05$, $P = 0.03$; *Ppp1cc*: $F_{3,11} = 3.57$, $P = 0.05$) and +1 (*Egr1*: $F_{3,11} = 3.59$, $P = 0.046$; *BdnfIV*: $F_{3,11} = 4.17$, $P = 0.03$; *Ppp1cc*: $F_{3,11} = 3.56$, $P = 0.05$) nucleosomes (Extended Data Fig. 8). Although we cannot conclude with certainty that H2A.Z binding at these loci is repressive, reduced H2A.Zac binding suggests that an activity-associated modification[14,16,23] is removed during systems consolidation in the cortex.

Next, we infused H2A.Z AAV into the medial pre-frontal cortex (mPFC; Fig. 4a) and confirmed a reduction in *H2afz* mRNA, as well as a 68.34% reduction in protein levels (Fig. 4b, c). H2A.Z depletion did not affect fear memory at the hippocampus-dependent 24 h time point, whereas significantly higher freezing was observed in H2A.Z-depleted mice at the two remote time points (30 days: $t_{14} = -5.28$, $P < 0.0001$; and 7 days: $t_8 = -3.07$, $P = 0.02$) (Fig. 4d–f). In separate mice, H2A.Z knockdown enhanced *Fos* expression irrespective of training 30 min after fear conditioning (main effect of virus: $F_{1,12} = 4.77$, $P = 0.049$) and reduced the expression of *Ppp3ca* (training–virus interaction: $F_{1,12} = 16.28$, $P < 0.002$) in untrained H2A.Z knockdown compared to untrained scrambled mice. The expression of remaining genes increased only as a function of fear conditioning (*Npas4*: $F_{1,12} = 108.65$, $P < 0.001$; *Arc*: $F_{1,12} = 156.00$, $P < 0.001$; *Egr1*: $F_{1,12} = 10.73$, $P = 0.007$; *Egr2*: $F_{1,12} = 115.03$, $P < 0.001$, *Fos*: $F_{1,12} = 63.77$, $P < 0.001$) (Fig. 4h–o). Although the virus was present throughout the consolidation and maintenance stages, the emergence of memory enhancement at 7 days, and altered *Fos* and *Ppp3ca* expression at 30 min, are consistent with an early role in systems consolidation.

Overall, H2A.Z has a restrictive effect on recent and remote memory. However, consistent with the wide genomic distribution of H2A.Z[11], our sequencing data demonstrate that its depletion both up- and downregulates 451 different genes, making it difficult to ascertain specific targets through which H2A.Z regulates memory.

While our data clearly indicate that H2A.Z is dynamically regulated during learning and memory, the basis for H2A.Z regulation in the central nervous system is not known and indeed, its regulation is not fully understood in any biological system. Recent studies identified DNA

methylation, sirtuin 1 and H3 acetylation at lysine 56[24–26] as negative regulators of H2A.Z. These factors have a known role in memory[3,7,27] and represent potential regulators of H2A.Z in fear conditioning. Notably, many changes in H2A.Z binding were not specific to associative learning, although H2A.Z exchange specificity was evident at numerous genes in the cortex and a subset of genes in CA1. Thus, although H2A.Z has the capacity for specific regulation of associative learning, its exchange is also sensitive to broader environmental stimuli.

Overall, we show that H2A.Z is a novel regulator of memory and introduce histone variant exchange as an additional epigenetic contributor to the complex coordination of gene expression in memory. Further, our data suggest that H2A.Z antagonists may provide a novel therapeutic target for memory disorders.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 12 May; accepted 23 July 2014.
Published online 14 September 2014.

1. Wang, S. H. & Morris, R. G. Hippocampal-neocortical interactions in memory formation, consolidation, and reconsolidation. *Annu. Rev. Psychol.* **61,** 49–79 (2010).
2. Lesburgueres, E. *et al.* Early tagging of cortical networks is required for the formation of enduring associative memory. *Science* **331,** 924–928 (2011).
3. Levenson, J. M. *et al.* Regulation of histone acetylation during memory formation in the hippocampus. *J. Biol. Chem.* **279,** 40545–40559 (2004).
4. Lubin, F. D., Roth, T. L. & Sweatt, J. D. Epigenetic regulation of *bdnf* gene transcription in the consolidation of fear memory. *J. Neurosci.* **28,** 10576–10586 (2008).
5. Miller, C. A., Campbell, S. L. & Sweatt, J. D. DNA methylation and histone acetylation work in concert to regulate memory formation and synaptic plasticity. *Neurobiol. Learn. Mem.* **89,** 599–603 (2008).
6. Miller, C. A. *et al.* Cortical DNA methylation maintains remote memory. *Nature Neurosci.* **13,** 664–666 (2010).
7. Miller, C. A. & Sweatt, J. D. Covalent modification of DNA regulates memory formation. *Neuron* **53,** 857–869 (2007).
8. Pina, B. & Suau, P. Changes in histones H2A and H3 variant composition in differentiating and mature rat brain cortical neurons. *Dev. Biol.* **123,** 51–58 (1987).
9. Santoro, S. W. & Dulac, C. The activity-dependent histone variant H2BE modulates the life span of olfactory neurons. *Elife* **1,** e00070 (2012).
10. Michod, D. *et al.* Calcium-dependent dephosphorylation of the histone chaperone DAXX regulates H3.3 loading and transcription upon neuronal activation. *Neuron* **74,** 122–135 (2012).
11. Bargaje, R. *et al.* Proximity of H2A.Z containing nucleosome to the transcription start site influences gene expression levels in the mammalian liver and brain. *Nucleic Acids Res.* **40,** 8965–8978 (2012).
12. Schauer, T. *et al.* CAST-ChIP maps cell-type-specific chromatin states in the *Drosophila* central nervous system. *Cell Rep* **5,** 271–282 (2013).
13. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Rev. Genet.* **13,** 484–492 (2012).
14. Bonisch, C. & Hake, S. B. Histone H2A variants in nucleosomes and chromatin: more or less stable? *Nucleic Acids Res.* **40,** 10719–10741 (2012).
15. Weber, C. M., Ramachandran, S. & Henikoff, S. Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. *Mol. Cell* **53,** 819–830 (2014).
16. Bellucci, L., Dalvai, M., Kocanova, S., Moutahir, F. & Bystricky, K. Activation of p21 by HDAC inhibitors requires acetylation of H2A.Z. *PLoS ONE* **8,** e54102 (2013).
17. Valdes-Mora, F. *et al.* Acetylation of H2A.Z is a key epigenetic modification associated with gene deregulation and epigenetic remodeling in cancer. *Genome Res.* **22,** 307–321 (2012).
18. Hardy, S. *et al.* The euchromatic and heterochromatic landscapes are shaped by antagonizing effects of transcription on H2A.Z deposition. *PLoS Genet.* **5,** e1000687 (2009).
19. Gevry, N., Chan, H. M., Laflamme, L., Livingston, D. M. & Gaudreau, L. p21 transcription is regulated by differential localization of histone H2A.Z. *Genes Dev.* **21,** 1869–1881 (2007).
20. Chauhan, S. & Boyd, D. D. Regulation of u-PAR gene expression by H2A.Z is modulated by the MEK-ERK/AP-1 pathway. *Nucleic Acids Res.* **40,** 600–613 (2012).
21. Nock, A., Ascano, J. M., Barrero, M. J. & Malik, S. Mediator-regulated transcription through the +1 nucleosome. *Mol. Cell* **48,** 837–848 (2012).
22. Smith, A. P. *et al.* Histone H2A.Z regulates the expression of several classes of phosphate starvation response genes but not as a transcriptional activator. *Plant Physiol.* **152,** 217–225 (2010).
23. Millar, C. B., Xu, F., Zhang, K. & Grunstein, M. Acetylation of H2AZ Lys 14 is associated with genome-wide gene activity in yeast. *Genes Dev.* **20,** 711–722 (2006).
24. Watanabe, S., Radman-Livaja, M., Rando, O. J. & Peterson, C. L. A histone acetylation switch regulates H2A.Z deposition by the SWR-C remodeling enzyme. *Science* **340,** 195–199 (2013).
25. Conerly, M. L. *et al.* Changes in H2A.Z occupancy and DNA methylation during B-cell lymphomagenesis. *Genome Res.* **20,** 1383–1390 (2010).
26. Baptista, T. *et al.* Regulation of histone H2A.Z expression is mediated by sirtuin 1 in prostate cancer. *Oncotarget* **4,** 1673–1685 (2013).
27. Gao, J. *et al.* A novel pathway regulates memory and plasticity via SIRT1 and miR-134. *Nature* **466,** 1105–1109 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements The authors' work is supported by DARPA grant HR0011-12-1-0015 and NIH grants MH091122, MH57014 (J.D.S.) and NSERC-PDF grant PDF 387473-10 (I.B.Z.). We would like to thank F. Sultan for providing RNA primers and K. Alison Margolies for providing the immunohistochemistry images.

Author Contributions J.D.S. and I.B.Z. conceived the experiments. I.B.Z. conducted the experiments and B.S.P. and D.M.E. assisted in performing the experiments. J.J.D. analysed the next-generation sequencing data.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. The next-generation sequencing data have been deposited in the GEO database and can be accessed using accession number GSE58797. Correspondence and requests for materials should be addressed to J.D.S. (dsweatt@uab.edu).

LETTER

doi:10.1038/nature13722

Epigenetic reprogramming that prevents transgenerational inheritance of the vernalized state

Pedro Crevillén[1]†, Hongchun Yang[1], Xia Cui[2], Christiaan Greeff[1]†, Martin Trick[1], Qi Qiu[2], Xiaofeng Cao[2] & Caroline Dean[1]

The reprogramming of epigenetic states in gametes and embryos is essential for correct development in plants and mammals[1]. In plants, the germ line arises from somatic tissues of the flower, necessitating the erasure of chromatin modifications that have accumulated at specific loci during development or in response to external stimuli. If this process occurs inefficiently, it can lead to epigenetic states being inherited from one generation to the next[2–4]. However, in most cases, accumulated epigenetic modifications are efficiently erased before the next generation. An important example of epigenetic reprogramming in plants is the resetting of the expression of the floral repressor locus *FLC* in *Arabidopsis thaliana*. *FLC* is epigenetically silenced by prolonged cold in a process called vernalization. However, the locus is reactivated before the completion of seed development, ensuring the requirement for vernalization in every generation. In contrast to our detailed understanding of the polycomb-mediated epigenetic silencing induced by vernalization, little is known about the mechanism involved in the reactivation of *FLC*. Here we show that a hypomorphic mutation in the jumonji-domain-containing protein ELF6 impaired the reactivation of *FLC* in reproductive tissues, leading to the inheritance of a partially vernalized state. ELF6 has H3K27me3 demethylase activity, and the mutation reduced this enzymatic activity *in planta*. Consistent with this, in the next generation of mutant plants, H3K27me3 levels at the *FLC* locus stayed higher, and *FLC* expression remained lower, than in the wild type. Our data reveal an ancient role for H3K27 demethylation in the reprogramming of epigenetic states in plant and mammalian embryos[5–7].

Many *A. thaliana* accessions overwinter before flowering, as a result of FRIGIDA (FRI)-mediated high-level expression of a floral repressor called FLC[8,9]. Prolonged cold during the weeks of winter antagonizes this activation and progressively epigenetically silences *FLC*. This process enables other floral promotion signals, such as day length, to induce flowering in spring. The epigenetic silencing of *FLC* involves polycomb-mediated chromatin regulation[10–12] and is maintained until embryogenesis, when *FLC* expression is reset to ensure a requirement for vernalization in every generation[13,14]. The resetting of *FLC* expression occurs in the early globular embryo[13,14]; then, *FLC* expression increases throughout embryo development until it reaches maximum levels when the seed has completely formed[14]. However, the molecular mechanisms underlying *FLC* resetting are unknown, and several factors that are required for the upregulation of *FLC* in vegetative tissues have been shown to be dispensable for *FLC* expression in the embryo[14]. One exception is the yeast SWR1 homologue PIE1 (ref. 14), although it is unclear whether *pie1* mutations are resetting-specific defects because these mutations strongly reduce *FLC* expression across the plant independently of vernalization status.

To dissect this resetting mechanism, we isolated mutants that are defective in the reactivation of *FLC* after vernalization (Extended Data Fig. 1a). The parental line was an *A. thaliana* Landsberg erecta (Ler) plant carrying an *FLC*::luciferase (*FLC*::*LUC*) translational fusion and an active *FRI* transgene[15]. We searched for plants in which *FLC* expression was silenced by vernalization but, unlike in the wild type, was not fully restored in the following generation (Fig. 1a), leading to inheritance of the vernalized state. The frequency of these mutations was low (with only 2 mutants identified from the progeny of 6,000 mutagenized parent lines), in contrast to the more common class of mutations, which involved early flowering before vernalization as a result of reduced *FLC* expression (Extended Data Fig. 1b). The first resetting mutant that was isolated was found to be recessive (Extended Data Fig. 2a) and flowered slightly earlier without vernalization than did the wild type (Fig. 1b). In the generation after vernalization, the mutant flowered even earlier and had significantly

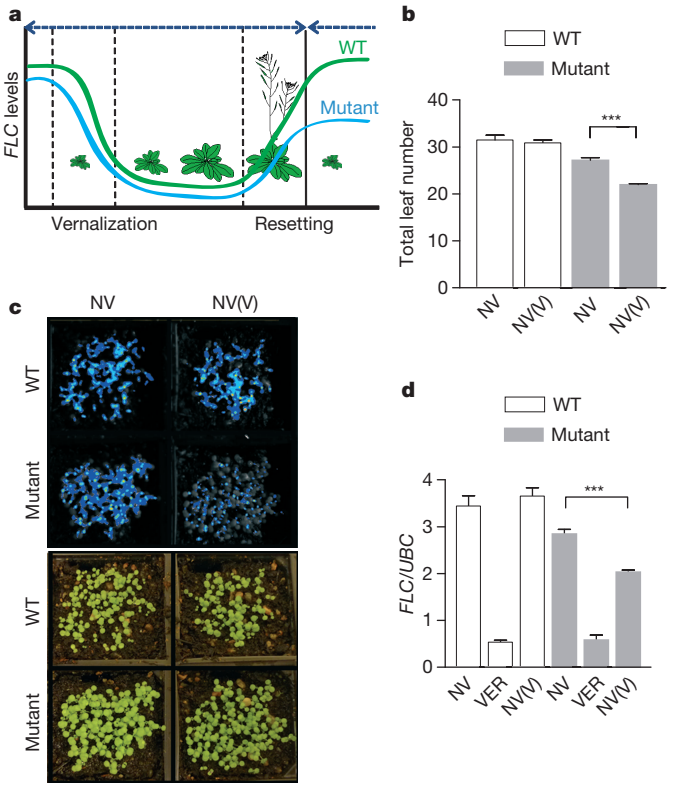


Figure 1 | Isolation and characterization of the resetting mutant. **a**, The rationale for the genetic screen. The parental wild type (WT) is Ler (*FRI*) *FLC*::*LUC*; the mutant is a resetting mutant. **b**–**d**, The resetting mutant is early flowering (**b**), with fewer leaves when bolting (with flowering time assayed as total leaf number), and maintains low *FLC* expression, as shown by *FLC*–luciferase imaging of 8-day-old seedlings (**c**) and by quantitative reverse transcription PCR (qRT–PCR) analysis normalized to *UBC* levels (**d**). Pseudocolour bioluminescent images (**c**, top) from blue (least intense) to red (most intense) and normal images (**c**, bottom) of the same plants are presented. The data are presented as the mean + s.e.m., $n = 20$ (**b**) and $n = 3$ (**d**); $^{***}P < 0.001$. NV, non-vernalized; NV(V), non-vernalized following vernalization in the previous generation; VER, vernalized.

[1]Department of Cell & Developmental Biology, John Innes Centre, Norwich Research Park, Norwich NR4 7UH, UK. [2]State Key Laboratory of Plant Genomics and National Center for Plant Gene Research, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences. Beijing 100101, China. †Present addresses: Centro de Biotecnología y Genómica de Plantas, UPM-INIA, 28223 Madrid, Spain (P.C.); Department of Biology, Copenhagen University, DK-2200 Copenhagen, Denmark (C.G.).

reduced *FLC* expression (Fig. 1c, d), albeit about fourfold higher than in fully vernalized seedlings (Fig. 1d). The resetting mutant therefore causes transgenerational inheritance of a partially vernalized state. The early flowering phenotype was stable for at least three generations following vernalization (Extended Data Fig. 2b) but was not enhanced by a second vernalization treatment in the later generations. No other strong developmental phenotypes were observed.

The mutant phenotype was strongly affected by the segregation of modifiers in a traditional Ler × Columbia (Col) cross, which is normally used for genetic mapping, and the mutation was only narrowed to a ~500 kilobase region on chromosome 5. We therefore sequenced the whole genome of the mutant plant and analysed the linkage of candidate single nucleotide polymorphisms (SNPs) in an F_2 population generated from a cross between the mutant and the isogenic progenitor line. This strategy identified a SNP in *ELF6* (AT5G04240) that co-segregated with the resetting phenotype (Extended Data Fig. 3). To confirm that the resetting phenotype was caused by this mutant allele (named *elf6-5*), we complemented the mutation using the wild-type *ELF6* gene under the control of its own regulatory sequences. Vernalized T_2 transgenic *elf6-5* lines carrying the wild-type *ELF6* transgene showed wild-type *FLC* expression levels in the siliques (Fig. 2a, b). Thus, we concluded that the single nucleotide mutation in *ELF6* causes the mutant phenotype.

ELF6 is a jumonji-C-domain-containing protein that is closely related to the histone H3 trimethylated lysine 27 (H3K27me3) demethylase REF6 (ref. 16), and it is expressed at low levels in seedlings but at high levels in flowers and embryos (Fig. 2c–f). In the *elf6*-5 mutants, an alanine is replaced with a valine (amino acid 424) at the carboxy-terminal end of the jumonji C domain (Fig. 2g). This amino acid is conserved in REF6 and the human H3K27me3 demethylases UTX (also known as KDM6A) and JMJD3 (also known as KDM6B) (Fig. 2g). This high degree of conservation suggests that this residue may be crucial for the function of the protein. A null *elf6* T-DNA insertion allele has been shown to be early flowering because of increased expression of *FT*[17], an integrator gene that promotes floral transition. In addition, we found less *FLC* expression in seedlings carrying this knockout allele (*elf6-3*) than in wild-type Col seedlings (Extended Data Fig. 4a), confirming that ELF6 regulates *FLC* expression. The early flowering phenotype and low *FLC* expression in *elf6-3* seedlings precluded observation of the resetting phenotype (Extended Data Fig. 4b). Although the different genetic backgrounds of the two alleles may complicate interpretation, these data suggest that the alanine-to-valine substitution in *elf6-5* plants confers a hypomorphic phenotype, affecting an activity that is particularly important for resetting *FLC* expression during reproductive development.

Consistent with a role for ELF6 in regulating *FLC* resetting, the *elf6-5* allele had a much larger effect on *FLC* expression in flowers and siliques than in seedlings (Fig. 3a, b). To define more precisely when the *elf6-5* mutant disrupts *FLC* expression, we measured *FLC* messenger RNA levels at different stages of silique development[18], a proxy for *FLC* expression in the embryo[13,14]. Low *FLC* mRNA levels were detected in young siliques from the vernalized parental line (SQ16 and SQ17a) (Fig. 3c), and these levels increased as the silique matured (SQ17b1 and SQ17b2), reaching a maximum when the silique started to desiccate and the embryo became fully developed (SQ18). In the vernalized resetting mutant, *FLC* mRNA was detected in young, developing siliques, but it was not upregulated to wild-type levels at the later stages (Fig. 3c). Comparison of siblings differing only by an *FLC*::*GUS* (β-glucuronidase) reporter[19] showed that *FLC*::*GUS* expression was lower in the early globular embryo of *elf6-5* mutants than in the wild type (Fig. 3d). This finding suggests that ELF6 increases *FLC* expression as the embryo develops. There may be no clear mechanistic separation between reprogramming the epigenetic state and setting the *FLC* expression level.

The *FLC* locus has a complex transcriptional circuitry, including a set of antisense transcripts called *COOLAIR* that are induced during vernalization but are also expressed in the warm[20]. We wondered whether the *elf6-5* resetting mutant also affected *COOLAIR* expression. Surprisingly, no difference between the mutant and the wild type was found, and *COOLAIR* transcripts were upregulated normally in the mutant in developing siliques (Fig. 3e). Therefore, in contrast to mutants in which both *FLC* sense and total *COOLAIR* expression levels change coordinately (for example, *fri* mutants), the mutation in *elf6-5* plants uncouples *FLC* sense and antisense regulation.

Many other loci are epigenetically modified during *A. thaliana* gamete formation and embryo development[21,22]. We tested whether the *elf6-5* allele influences transposon expression, by analysing specific short interfering

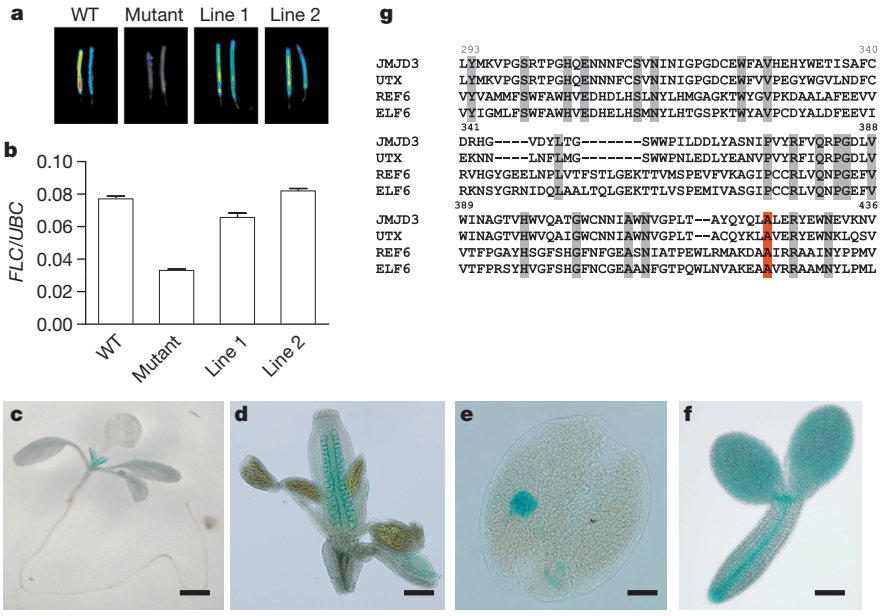


Figure 2 | Mapping of the resetting mutant. **a**, **b**, An *ELF6* genomic construct complements the resetting mutant. *FLC*–luciferase imaging (**a**) and *FLC* qRT–PCR data (**b**) from mature siliques from the vernalized WT, the *elf6-5* mutant and representative T_2 *elf6-5* (*pELF6*::*ELF6*) lines (Line1 and Line2). Pseudocolour bioluminescent image (**a**) from blue (least intense) to red (most intense). **c**–**f**, *ELF6*::*GUS* (blue) expression profile in a 7-day-old seedling (**c**), ovules (**d**), a globular embryo (**e**) and a mature embryo (**f**). Scale bars, 5 mm (**c**), 250 μm (**d**), 50 μm (**e**, **f**). **g**, The ELF6 amino acid residue that is mutated in *elf6-5* mutants is conserved (red; A, alanine). A sequence alignment of the jumonji C domain of *A. thaliana* ELF6 and REF6 and human JMJD3 and UTX proteins is shown. Highly conserved residues are shaded in grey. The numbering refers to the ELF6 amino acid position.

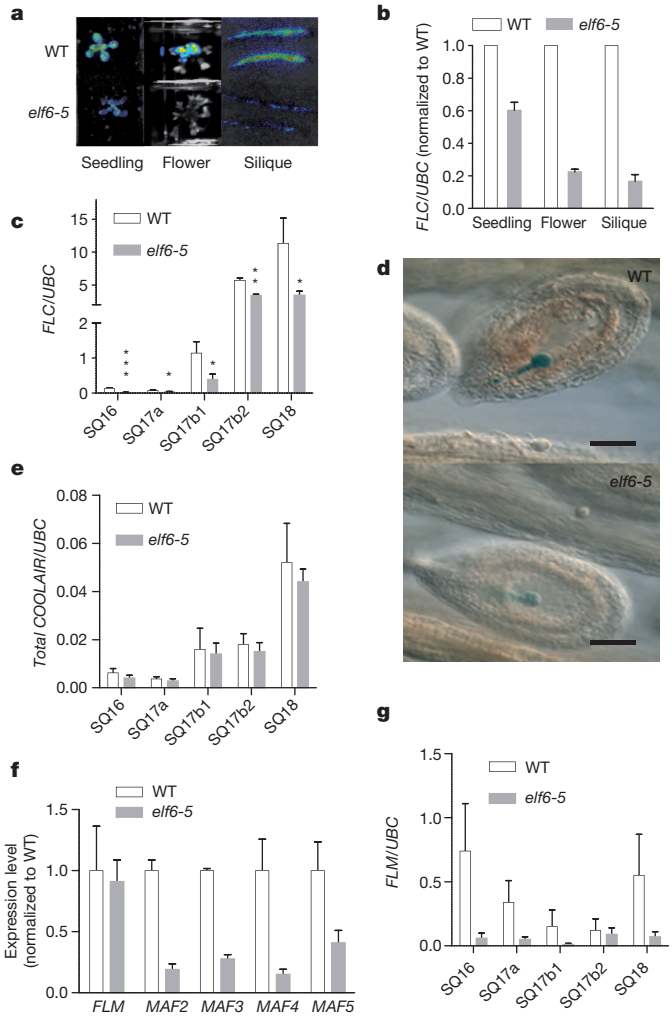


Figure 3 | Characterization of the *elf6-5* resetting mutant. **a**, **b**, *FLC*–luciferase imaging (**a**) and *FLC* qRT–PCR data (**b**) for tissues from the WT and the *elf6-5* mutant in the generation following vernalization. The data are presented as the mean + s.e.m., $n = 6$. Pseudocolour bioluminescent image (**a**) from blue (least intense) to red (most intense). **c**, qRT–PCR data for vernalized WT and *elf6-5* siliques[18]: immediately after fertilization, with petals still attached (SQ16); small and without petals (SQ17a); first (SQ17b1) and last (SQ17b2) mature green siliques; and yellow siliques (SQ18). The data are presented as the mean + s.e.m., $n = 4$; $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$ compared with WT. **d**, *FLC*::*GUS* (blue) expression in vernalized WT (top) and *elf6-5* (bottom) early globular embryos. Scale bars, 100 μm. **e**, qRT–PCR shows that *COOLAIR* levels are not affected in *elf6-5* siliques. The data are presented as the mean + s.e.m., $n = 5$. **f**, qRT–PCR data showing that the *MAF2*, *MAF3*, *MAF4* and *MAF5* genes are misregulated in *elf6-5* seedlings in the generation following vernalization. The data are presented as the mean + s.e.m., $n = 3$. **g**, *FLM* has reduced expression in vernalized *elf6-5* siliques. The data are presented as the mean + s.e.m., $n = 3$.

RNAs (siRNAs) produced during seed development[22], including an siRNA homologous to the flanking 3′ region of the *FLC* locus that accumulates preferentially in siliques[23]. Using sensitive northern blot analyses, we did not detect any difference in the amount of these siRNAs between vernalized siliques from *elf6-5* and the parental line (Extended Data Fig. 5). We then asked whether the *elf6-5* allele influenced the expression of other *FLC*-family members[24]. *MAF2*, *MAF3*, *MAF4* and *MAF5* were downregulated in *elf6-5* seedlings and were expressed below the detection level in siliques (Fig. 3f), whereas *FLM* (also known as *MAF1*) expression was unchanged in seedlings but strongly downregulated in *elf6-5* siliques (Fig. 3f, g). The vernalization response in winter-annual *A. thaliana* accessions (containing an active *FRI* allele) depends predominantly on FLC activity[24]; however, in the rapid-cycling Col (*fri*) genotype, all *MAF* genes appear to be direct targets of the polycomb machinery[24,25], indicating that the ELF6-regulatory mechanism that has been elaborated for *FLC* may have more general functions in the *A. thaliana* genome.

Since ELF6 is closely related to the H3K27me3 demethylase REF6, we tested the enzymatic activity of both wild-type and mutant ELF6, by using an *in vivo* histone demethylation assay[16]. We found that transient expression of *A. thaliana* ELF6 resulted in reduced H3K27me2 and H3K27me3 levels in tobacco (*Nicotiniana benthamiana*) leaves (Fig. 4a); no changes were found in the levels of H3K27me1, H3K4me3, H3K9me2 and H3K36me3 (Fig. 4a and Extended Data Fig. 6). These data show that *A. thaliana* ELF6 has histone demethylase activity specific for H3K27me2 and H3K27me3. The *elf6*-5 mutant carries an alanine-to-valine substitution at the C-terminal end of the jumonji C domain. The mutation

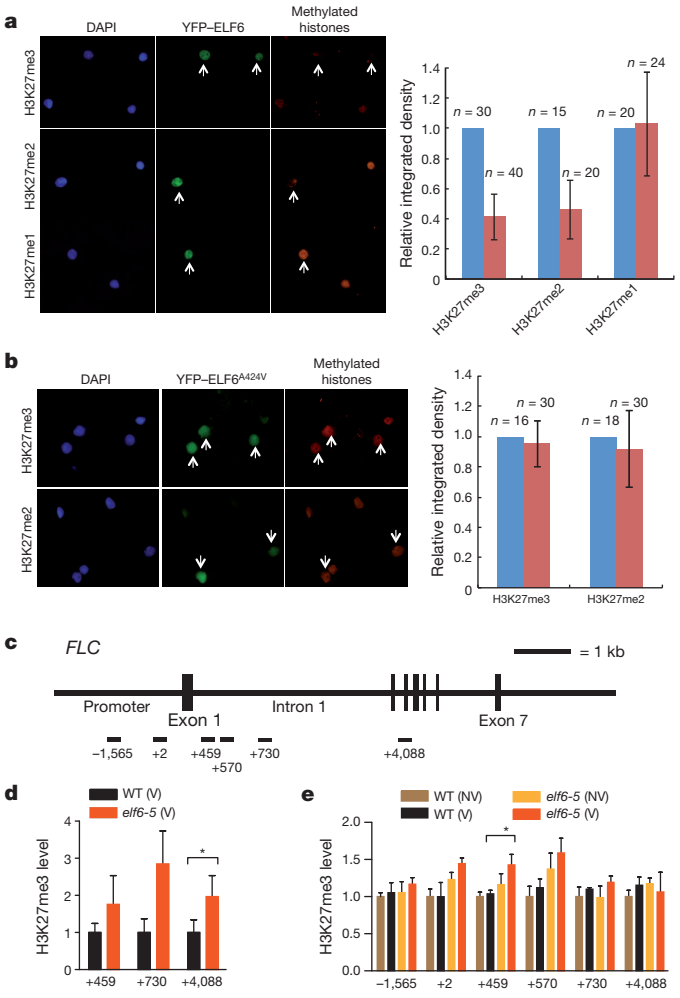


Figure 4 | ELF6 shows H3K27 histone demethylase activity. **a**, **b**, Overexpression of a yellow fluorescent protein (YFP)–ELF6 fusion protein reduces H3K27me3 and H3K27me2 levels but not H3K27me1 levels (**a**, left). Overexpression of YFP–ELF6^{A424V} has no effect on H3K27 methylation (**b**, left). Histone methylation was visualized by immunostaining (red). Nuclei (arrows) from transfected cells were visualized by the YFP signal (green) and stained with DAPI (4′,6-diamidino-2-phenylindole) (blue). The histograms quantify the methylation levels in the nuclei of transfected (red) and non-transfected (blue) cells. The data are presented as the mean ± s.d. **c**, *FLC* regions analysed in ChIP. kb, kilobase. **d**, H3K27me3 levels in *elf6-5* and WT siliques (stage SQ16–SQ17a) from vernalized plants. The data are presented as the mean + s.e.m., $n = 3$. The H3K27me3 level in WT (V) plants is significantly lower than that in *elf6-5* (V) in the *FLC* region +4,088 ($^{*}P < 0.05$). **e**, The H3K27me3 levels in progeny derived from parents that had (V) or had not (NV) been vernalized. The H3K27me3 level in WT (V) plants is significantly lower than that in *elf6-5* (V) in the *FLC* region +459 ($^{*}P < 0.05$). The data are presented as the mean + s.e.m., $n = 3$.

of this highly conserved residue (Fig. 2g) reduced H3K27 demethylase activity in our assay (Fig. 4b). To test whether this reduced activity influenced H3K27me3 levels at the *FLC* locus *in vivo*, we performed chromatin immunoprecipitation (ChIP) experiments. In wild-type Ler (*FRI*) plants, H3K27me3 levels increased by about twofold to fourfold in vernalized seedlings but were reduced to almost non-vernalized levels in vernalized siliques (Extended Data Fig. 7). When the resetting mutant was analysed, we found that the H3K27me3 levels were higher at *FLC* in vernalized young siliques of *elf6-5* mutants than of the parental line (Fig. 4d). ChIP analysis on seedlings of the generation following vernalization also showed increased levels of H3K27me3 over various regions of *FLC* in *elf6-5* mutants (Fig. 4e). These experiments were performed using whole seedlings or siliques; therefore, the data should be interpreted with caution because they are derived from a mixture of tissues. The vernalization-independent increase in H3K27me3 levels in *elf6-5* mutants and the phenotype of the *elf6* loss-of-function mutant make it likely that ELF6 has broader functions than simply resetting H3K27 methylation after vernalization. Nevertheless, all of these data are consistent with the reduced *FLC* expression during embryo development in *elf6-5* mutants involving perturbed H3K27me3 dynamics that affect *FLC* resetting and result in the inheritance of a partially vernalized state.

This effective impairment of the reduction of H3K27me3 levels at *FLC* leads to transgenerational inheritance of a partially vernalized state (Fig. 1a–d). In nature, the consequences of this impairment would be to misalign the developmental program of the plant with respect to the environmental conditions. The sensitivity of *FLC* resetting to the reduced function *elf6-5* allele may indicate that the requirement for H3K27me3 demethylase activity is highest at this post-vernalization stage of development, potentially explaining the differences in phenotype between *elf6-5* and the null allele *elf6-3*. Functional redundancy between the three close homologues REF6, ELF6 and JMJ13 (AT5G46910) may also vary throughout development[16]. It will be interesting to see whether histone variants that are known to change in expression during embryogenesis[26] are also involved in *FLC* resetting. In most eukaryotic genomes, a large proportion of chromatin is decorated with H3K27me3, probably explaining why the erasure of these methyl groups is a tightly controlled event during development and germ cell formation[27]. Further characterization of the *FLC* resetting process should provide greater insight into the molecular mechanism underlying genome reprogramming in eukaryotic organisms.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 5 November 2013; accepted 29 July 2014.
Published online 14 September 2014.

1. Feng, S., Jacobsen, S. E. & Reik, W. Epigenetic reprogramming in plant and animal development. *Science* **330,** 622–627 (2010).
2. Paszkowski, J. & Grossniklaus, U. Selected aspects of transgenerational epigenetic inheritance and resetting in plants. *Curr. Opin. Plant Biol.* **14,** 195–203 (2011).
3. Becker, C. *et al.* Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature* **480,** 245–249 (2011).
4. Schmitz, R. J. *et al.* Transgenerational epigenetic instability is a source of novel methylation variants. *Science* **334,** 369–373 (2011).
5. Mansour, A. A. *et al.* The H3K27 demethylase Utx regulates somatic and germ cell epigenetic reprogramming. *Nature* **488,** 409–413 (2012).
6. Canovas, S., Cibelli, J. B. & Ross, P. J. Jumonji domain-containing protein 3 regulates histone 3 lysine 27 methylation during bovine preimplantation development. *Proc. Natl Acad. Sci. USA* **109,** 2400–2405 (2012).
7. Zhao, W. *et al.* Jmjd3 inhibits reprogramming by upregulating expression of INK4a/Arf and targeting PHF20 for ubiquitination. *Cell* **152,** 1037–1050 (2013).
8. Johanson, U. Molecular analysis of *FRIGIDA*, a major determinant of natural variation in *Arabidopsis* flowering time. *Science* **290,** 344–347 (2000).
9. Sheldon, C. C. *et al.* The *FLF MADS* box gene: a repressor of flowering in *Arabidopsis* regulated by vernalization and methylation. *Plant Cell* **11,** 445–458 (1999).
10. Gendall, A. R., Levy, Y. Y., Wilson, A. & Dean, C. The *VERNALIZATION 2* gene mediates the epigenetic regulation of vernalization in *Arabidopsis. Cell* **107,** 525–535 (2001).
11. Song, J., Angel, A., Howard, M. & Dean, C. Vernalization: a cold-induced epigenetic switch. *J. Cell Sci.* **125,** 3723–3731 (2012).
12. De Lucia, F., Crevillen, P., Jones, A. M. E., Greb, T. & Dean, C. A PHD–polycomb repressive complex 2 triggers the epigenetic silencing of *FLC* during vernalization. *Proc. Natl Acad. Sci. USA* **105,** 16831–16836 (2008).
13. Sheldon, C. C. *et al.* Resetting of *FLOWERING LOCUS C* expression after epigenetic repression by vernalization. *Proc. Natl Acad. Sci. USA* **105,** 2214–2219 (2008).
14. Choi, J. *et al.* Resetting and regulation of *FLOWERING LOCUS C* expression during *Arabidopsis* reproductive development. *Plant J.* **57,** 918–931 (2009).
15. Mylne, J., Greb, T., Lister, C. & Dean, C. Epigenetic regulation in the control of flowering. *Cold Spring Harb. Symp. Quant. Biol.* **69,** 457–464 (2004).
16. Lu, F., Cui, X., Zhang, S., Jenuwein, T. & Cao, X. *Arabidopsis* REF6 is a histone H3 lysine 27 demethylase. *Nature Genet.* **43,** 715–719 (2011).
17. Noh, B. *et al.* Divergent roles of a pair of homologous jumonji/zinc-finger-class transcription factor proteins in the regulation of *Arabidopsis* flowering time. *Plant Cell* **16,** 2601–2613 (2004).
18. Roeder, A. H. K. & Yanofsky, M. F. Fruit development in *Arabidopsis. Arabidopsis Book* **4,** e0075 (2006).
19. Bastow, R. *et al.* Vernalization requires epigenetic silencing of *FLC* by histone methylation. *Nature* **427,** 164–167 (2004).
20. Ietswaart, R., Wu, Z. & Dean, C. Flowering time control: another window to the connection between antisense RNA and chromatin. *Trends Genet.* **28,** 445–453 (2012).
21. Slotkin, R. K. *et al.* Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell* **136,** 461–472 (2009).
22. Mosher, R. A. *et al.* Uniparental expression of PolIV-dependent siRNAs in developing endosperm of *Arabidopsis. Nature* **460,** 283–286 (2009).
23. Swiezewski, S. *et al.* Small RNA-mediated chromatin silencing directed to the 3′ region of the *Arabidopsis* gene encoding the developmental regulator, FLC. *Proc. Natl Acad. Sci. USA* **104,** 3633–3638 (2007).
24. Alexandre, C. M. & Hennig, L. *FLC* or not *FLC*: the other side of vernalization. *J. Exp. Bot.* **59,** 1127–1135 (2008).
25. Kim, D.-H. & Sung, S. Coordination of the vernalization response through a *VIN3* and *FLC* gene family regulatory network in *Arabidopsis. Plant Cell* **25,** 454–469 (2013).
26. Ingouff, M. *et al.* Zygotic resetting of the HISTONE 3 variant repertoire participates in epigenetic reprogramming in *Arabidopsis. Curr. Biol.* **20,** 2137–2143 (2010).
27. Surani, M. A., Hayashi, K. & Hajkova, P. Genetic and epigenetic regulators of pluripotency. *Cell* **128,** 747–762 (2007).

Acknowledgements We thank Dean laboratory members and A. Surani for discussions. The Dean laboratory is supported by the UK Biotechnology and Biological Sciences Research Council grants BB/G009562/1 and BB/C517633/1 and by a European Research Council Advanced Investigator grant (233039 ENVGENE). The Cao laboratory is supported by National Basic Research Program of China grants 2013CB967300 and 2011CB915400 and by the National Natural Science Foundation of China grant 31271363.

Author Contributions P.C. and C.D. designed the research. P.C., H.Y., X. Cui, C.G. and Q.Q. performed experiments. M.T. conducted deep sequencing data analysis. X. Cao contributed new reagents and analytical tools. P.C., H.Y. and C.D. analysed the data and wrote the paper. All authors discussed the results and commented on the manuscript.

Author Information Genomic DNA deep sequencing data for the parental Ler-derived plant and the resetting mutant line have been deposited in the European Nucleotide Archive database under accession number PRJEB6498. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.D. (caroline.dean@jic.ac.uk).

LETTER

doi:10.1038/nature13709

A structure–based mechanism for tRNA and retroviral RNA remodelling during primer annealing

Sarah B. Miller[1]†*, F. Zehra Yildiz[1]*, Jennifer A. Lo[1], Bo Wang[1] & Victoria M. D'Souza[1]

To prime reverse transcription, retroviruses require annealing of a transfer RNA molecule to the U5 primer binding site (U5-PBS) region of the viral genome[1,2]. The residues essential for primer annealing are initially locked in intramolecular interactions[3–5]; hence, annealing requires the chaperone activity of the retroviral nucleocapsid (NC) protein to facilitate structural rearrangements[6]. Here we show that, unlike classical chaperones, the Moloney murine leukaemia virus NC uses a unique mechanism for remodelling: it specifically targets multiple structured regions in both the U5-PBS and tRNA^{Pro} primer that otherwise sequester residues necessary for annealing. This high-specificity and high-affinity binding by NC consequently liberates these sequestered residues—which are exactly complementary—for intermolecular interactions. Furthermore, NC utilizes a step-wise, entropy-driven mechanism to trigger both residue-specific destabilization and residue-specific release. Our structures of NC bound to U5-PBS and tRNA^{Pro} reveal the structure-based mechanism for retroviral primer annealing and provide insights as to how ATP-independent chaperones can target specific RNAs amidst the cellular milieu of non-target RNAs.

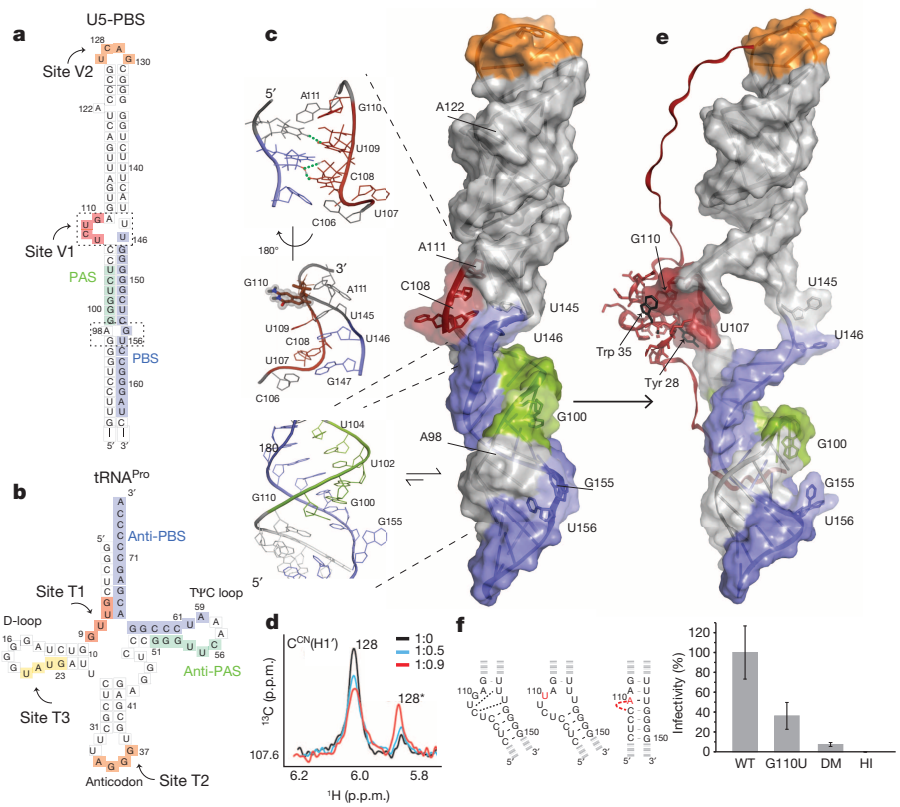


Figure 1 | Structure, function and NC interaction of U5-PBS. **a**, **b**, Secondary structures of U5-PBS (**a**) and tRNA^{Pro} (**b**) with complementary PBS/anti-PBS and PAS/anti-PAS sequences shown in blue and green, respectively. The two internal loops in U5-PBS are boxed, and the first, second and third high-affinity NC binding site sequences are shown in red, orange and yellow, respectively. The tRNA^{Pro} residue numbering reflects the canonical numbering scheme for all tRNAs and, hence, the number 17 is omitted. **c**, NMR structure of the free U5-PBS RNA. Insets: top, zoom-in view of the sequestration of U_{145} and $^{+1}U_{146}$ residues by U_{109} and C_{108} of the $UCUG_{110}$ NC binding site; middle, G_{110} aromatic ring is turned outside of the helix in a conformation poised for NC interaction; bottom, the minor, extruded conformation of G_{155} is shown. PAS residues are sequestered by alternating G–C and G–U base pairs in both conformations. **d**, One-dimensional slice of a ^{1}H–^{13}C spectrum depicting the NC tail-mediated increase of the minor conformation (indicated by an asterisk) of the tetraloop C_{128} residue. On the basis of the population distribution in the pre-bound state, we estimate the extruded conformation to have a free energy of ~1.4 kcal mol^{-1} greater than the stacked conformation. **e**, Structure of NC bound to $UCUG_{110}$ via the zinc finger (the Trp 35 and Tyr 28 stacking interactions are in black) showing that the protein tails can extend to contact the $UCAG_{130}$ tetraloop and residues in and near the ($G_{97}A$, $G_{155}U$) internal loop. **f**, Left, secondary structures of the wild-type (WT), G110U and deletion mutant (DM). Right, reduction of viral infectivity is observed in mutants ($n = 6$). Error bars indicate standard deviations ($n = 6$ for both packaging and infectivity experiments). As a negative control, heat-inactivated (HI) virions were used for infection. A packaging assay was also done to confirm that the mutations do not affect genome encapsidation (see Extended Data Fig. 4g).

[1]Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA. †Present address: Department of Biology, Georgetown University, Washington DC 20057, USA.
*These authors contributed equally to this work.

Retroviruses preferentially use specific host tRNAs as primers for the first step of reverse transcription; for example, human immunodeficiency virus requires $tRNA^{Lys\,3}$, while Moloney murine leukaemia virus (MLV) uses $tRNA^{Pro\,1,2}$. Two distinct sequences in the tRNA anneal to complementary sequences in the retroviral U5-PBS domain to form the initiation complex: the 3′ end of the tRNA acceptor stem anneals to the 18-nucleotide PBS sequence[7], while a portion of the tRNA TΨC arm base-pairs with a primer activation signal (PAS)[8,9] (Fig. 1a, b). However, for primer annealing to occur, favourable intramolecular associations involving both the PBS and PAS in U5-PBS and the complementary anti-PBS and anti-PAS sequences in tRNA must first be disrupted by NC chaperone proteins. Mechanistically, all NC proteins have thus far been thought to function as classical, ATP-independent chaperones, using both their zinc-finger domain(s) and unstructured tails for this process[10,11] (Extended Data Fig. 1a). ATP-independent chaperones are known to permit RNA molecules to access higher energy conformations and then allow refolding by rapidly dissociating from the RNA during the process[12]. These transitory, low-affinity interactions generally necessitate the coating of an RNA with many molecules of chaperone[12] (Extended Data Fig. 1b). In addition, and in contrast, to the transient interactions of NC proteins[13–15], the NC zinc fingers are also capable of sequence-specific, high-affinity binding to RNA[16]. However, this mode of interaction has, until now, been thought to be used exclusively for the recognition of the genome via interaction with the Ψ-genome packaging signal during viral assembly (Supplementary Discussion 1). To gain insights into the mechanism of NC-mediated primer annealing in MLV, a prototypical retrovirus, we solved structures of both the genomic U5-PBS RNA and the $tRNA^{Pro}$ primer, both in the free form and in complex with MLV NC proteins, by NMR spectroscopy.

The free U5-PBS is a largely linear molecule capped by a structured tetraloop ($UCAG_{130}$) and contains one single-nucleotide bulge (A_{122}) and two internal loops (($UCUGA_{111}$, UU_{146}) and (GA_{98}, GU_{156})) (Fig. 1a, c, Extended Data Figs 2, 3 and Extended Data Table 1). In the absence of NC, the ($UCUGA_{111}$, UU_{146}) internal loop maintains a distinct, folded configuration in which residue C_{108} of the NC binding site sequesters the 5′ end of the PBS sequence ($^{+1}U_{146}$) via an intramolecular ribose zipper interaction[17,18] (superscript denotes the PBS position from 5′ to 3′) (Fig. 1c and Extended Data Fig. 3e–g). Residue U_{109} of the NC binding site also base pairs with U_{145}, which is the first template residue read by reverse transcriptase. Continuous intramolecular base stacking interactions from U_{144} to $^{+2}G_{147}$ further serve to tether the 5′ end of the PBS inside the internal loop. On the other side of the bulge, continuous nuclear Overhauser enhancements (NOEs) indicate the stacking of C_{106} with C_{108} and, hence, extrusion of residue U_{107}. Importantly, residue G_{110} exhibits a *syn* glycosidic torsion angle and faces the major groove, making it poised for interaction with NC (Fig. 1c). In the NC-bound structure, the NC zinc finger binds the $UCUG_{110}$ sequence of the internal loop (dissociation constant (K_d) = 33 ± 3 nM; Extended Data Fig. 4a) in a mode similar to that previously described for $UCUG_{309}$ in the MLV Ψ-genome packaging signal[19–21] (see Supplementary Discussion 1 and 2). Notably, since NC-binding residues are initially involved in sequestering the first template (U_{145}) and the first PBS ($^{+1}U_{146}$) residues, the well-defined internal loop structure is mutually exclusive with NC binding. Thus, NC binding liberates U_{145} and $^{+1}U_{146}$ for primer annealing and reverse transcription initiation (Fig. 1e).

Whereas the NC tails are not involved in binding the Ψ-packaging signal, in U5-PBS, the NC tails specifically remodel the (GA_{98}, GU_{156}) internal loop and the $UCAG_{130}$ capping tetraloop (Fig. 1e). In the absence of NC, the six PAS residues in the lower stem, $_{+1}G_{99}$ to $_{+6}U_{104}$ (subscript with plus sign denotes the PAS position from 5′ to 3′), form base pairs with PBS residues (Fig. 1c and Extended Data Fig. 2a), and hence are not available for primer annealing. The preceding (GA_{98}, GU_{156}) internal loop, however, exists in multiple conformations; in the major form, a continuous internal stacking of all residues leads to the formation of tandem A_{98}–$^{+10}G_{155}$ and G_{97}–$^{+11}U_{156}$ non-canonical base pairs, while in the minor, destabilized form, the $^{+10}G_{155}$ and $^{+11}U_{156}$ PBS residues are extra-helical (Fig. 1c and Extended Data Figs 2a, 3h–j). Similarly, the capping $UCAG_{130}$ tetraloop forms a YNMG-type structure[22] (Extended Data Fig. 3a), with the C_{128} base either stacked on A_{129} or, in a small population, extruded from the structure (Fig. 1d and Extended Data Fig. 3c). Interaction with the NC tails alters the equilibria between the two conformations in favour of the minor, destabilized conformations (Fig. 1d) and hence leads to release of the 10th and 11th ($^{+10}G_{155}$ and $^{+11}U_{156}$) PBS residues and the C_{128} tetraloop residue. The destabilization of the latter is important for cooperative binding of a second NC to the tetraloop (Extended Data Fig. 4a; see Supplementary Discussion 3). Thus, the positively charged NC tails do not globally destabilize U5-PBS but instead specifically target residues inherently predisposed for destabilization. Furthermore, because PAS residues immediately follow the destabilized internal loop (Fig. 1a), the NC tails also specifically perturb residues $_{+1}G_{99}$, $_{+2}G_{100}$, and $_{+3}G_{101}$ (Extended Data Fig. 3k). Interestingly, both NC tails (Ala 1–Arg 17 and Arg 4–Leu 56) remain disordered, indicating that destabilization must occur via transient interactions that are nevertheless residue-specific owing to the inherent accessibility of the particular RNA residues and the orientation constraints imposed by the zinc finger binding to $UCUG_{110}$ (Fig. 1e and Extended Data Fig. 3l). In live viruses, a G110U mutant designed to liberate the U_{145} and $^{+1}U_{146}$ residues and a deletion mutant designed to sequester them exhibited only

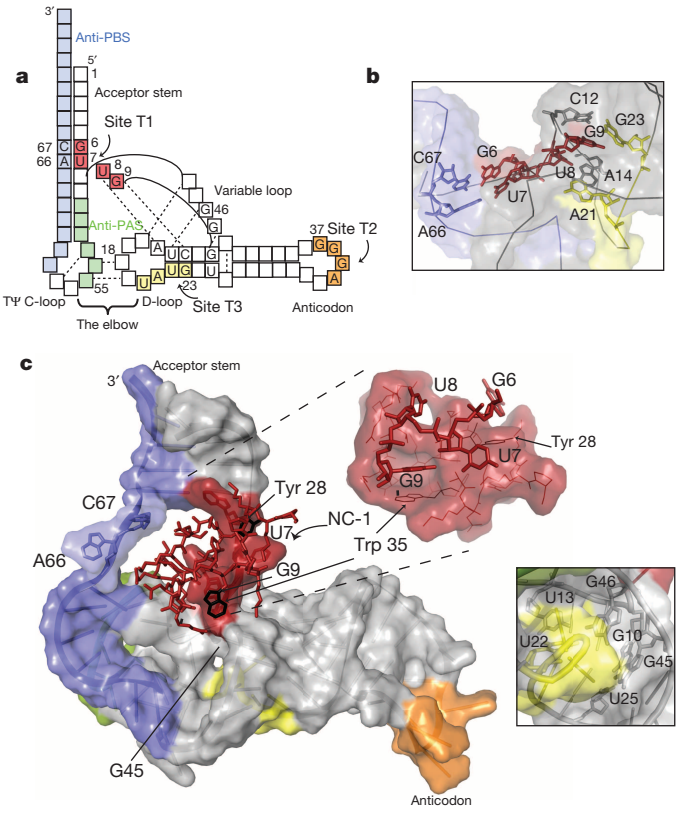


Figure 2 | Structure and NC interaction of site T1 in $tRNA^{Pro}$. **a**, Cartoon representation of the L-shaped $tRNA^{Pro}$ structure. Dashed lines show the long-range elbow and base triple interactions with the D-stem loop. Solid lines denote the covalent linkages. **b**, Three-dimensional model of the free $tRNA^{Pro}$ showing site T1 residues $GUUG_9$ (red) involved in intramolecular interactions. **c**, Structure of the NC–$tRNA^{Pro}$ complex. The aromatic NC residues Tyr 28 and Trp 35 are shown in black. The zinc finger interaction with the $GUUG_9$ binding site is shown in red. The tails are excluded for simplicity since they form random coils and do not specifically interact with the tRNA (Extended Data Fig. 9f). Top inset shows a close-up of the zinc finger interaction, with G_9 inserted into the hydrophobic pocket of the NC protein due to stacking interactions of Trp 35. Bottom inset shows the interaction of the variable loop with the core of the tRNA molecule that is maintained after interaction with NC-1.

56% and 7%, respectively, of wild-type MLV infectivity (Fig. 1f). While the severe infectivity defect of deletion mutant virions confirms the importance of high-affinity NC binding in releasing the 5′ end of the PBS, the partial defect observed for G110U virions indicates that tail-mediated interactions are also required for optimal function.

In $tRNA^{Pro}$, NC binding occurs first at $GUUG_9$ (site T1; 4 ± 2 nM) followed by the anticodon loop $AGGG_{37}$ (site T2; 13 ± 3 nM) and then the D-stem loop $UAUG_{23}$ (site T3; 834 ± 343 nM) (Extended Data Fig. 4b). Importantly, titration of MLV NC into the human immunodeficiency virus (HIV) primer, $tRNA^{Lys\,3}$, did not lead to NMR chemical shift perturbations, thus confirming the specificity of MLV NC for $tRNA^{Pro}$ (Extended Data Fig. 1f, g; for assignment strategy of $tRNA^{Pro}$ see Extended Data Figs 5–8 and Supplementary Discussion 4). The structure of the first NC bound to $tRNA^{Pro}$ shows the zinc finger making extensive contacts with the $GUUG_9$ sequence that links the acceptor stem with the D-stem loop, with G_9 stacking within the zinc finger pocket (Fig. 2, Extended Data Table 1 and Extended Data Fig. 9a). Importantly, before NC binding, all four $GUUG_9$ residues are involved in intramolecular interactions: G_6 and U_7 are part of the acceptor stem, and U_8 and G_9 are involved in core tertiary interactions (Fig. 2a, b and Extended Data Figs 6, 8a). Similar to U5-PBS, because these contacts are mutually exclusive with NC interactions, NC binding leads to major RNA remodelling events. First, since residues G_6 and U_7 are initially base paired with anti-PBS residues $^{-10}C_{67}$ and $^{-11}A_{66}$ (superscript with minus sign denotes the anti-PBS position from 3′ to 5′), respectively, NC binding releases these sequestered anti-PBS residues (Fig. 2c and Extended Data Fig. 9b, c). Second, because U_8 and G_9 are involved in core tertiary interactions via triple base formation with the D-stem loop (U_8:A_{14}–A_{21}, G_9:C_{12}–G_{23}) (Fig. 2a, b and Extended Data Fig. 8a), NC binding disrupts these core tertiary interactions (Extended Data Fig. 9c). As a result of this rearrangement, D-stem residues A_{21} and G_{23}, which are part of the $UAUG_{23}$ site T3 sequence, are made partially available for the third NC binding event. Globally, while the helical arrangement between the TΨC-stem and the acceptor stem is lost, the helical stacking between the D-stem and anticodon stem is preserved (Extended Data Fig. 9d, e), as is the variable loop interaction and the TΨC-loop:D-loop interface (the 'elbow') (Fig. 2a, c and Fig. 3a).

In comparison with sites T1 and T3 (see later), there is a marked mechanistic difference in the remodelling activity of the NC that binds the second site, T2—this NC uses its tails to achieve residue-specific destabilization. After the first NC binding event, the second NC accesses the residual elbow structure by anchoring its zinc finger to the distal $AGGG_{37}$ sequence in the anticodon loop, with the G_{37} base stacking inside the zinc finger (Fig. 3b and Extended Data Fig. 9g). While the elbow interaction is maintained, the NC tails specifically perturb D-loop and TΨC-loop residues G_{16} and A_{59}, respectively (Fig. 3c, d), which are in close proximity to each other (Fig. 3b). Prior to NC binding, these residues are extruded out of their respective loops and are hence available for NC interaction. Thus, as in the interaction with U5-PBS, the NC tails do not cause global destabilization of $tRNA^{Pro}$ but instead target specific, already accessible residues for remodelling.

We also structurally characterized the third NC binding event using the $tRNA^{Pro}$–$T1_MT2_M$ construct (see Supplementary Discussion 5). Our structures show that NC zinc finger binding to $UAUG_{23}$ in the D-stem disrupts the entire helix and, because the D-stem architecture is crucial for the D-loop–TΨC interaction, eliminates the residual elbow tertiary structure (Fig. 3a, e and Extended Data Fig. 9h, i). Consequently, the interactions between D-loop residues G_{18} and G_{19} and TΨC-loop anti-PAS residues $_{-1}C_{56}$ and $_{-2}U_{55}$ are lost, leading to the release of these

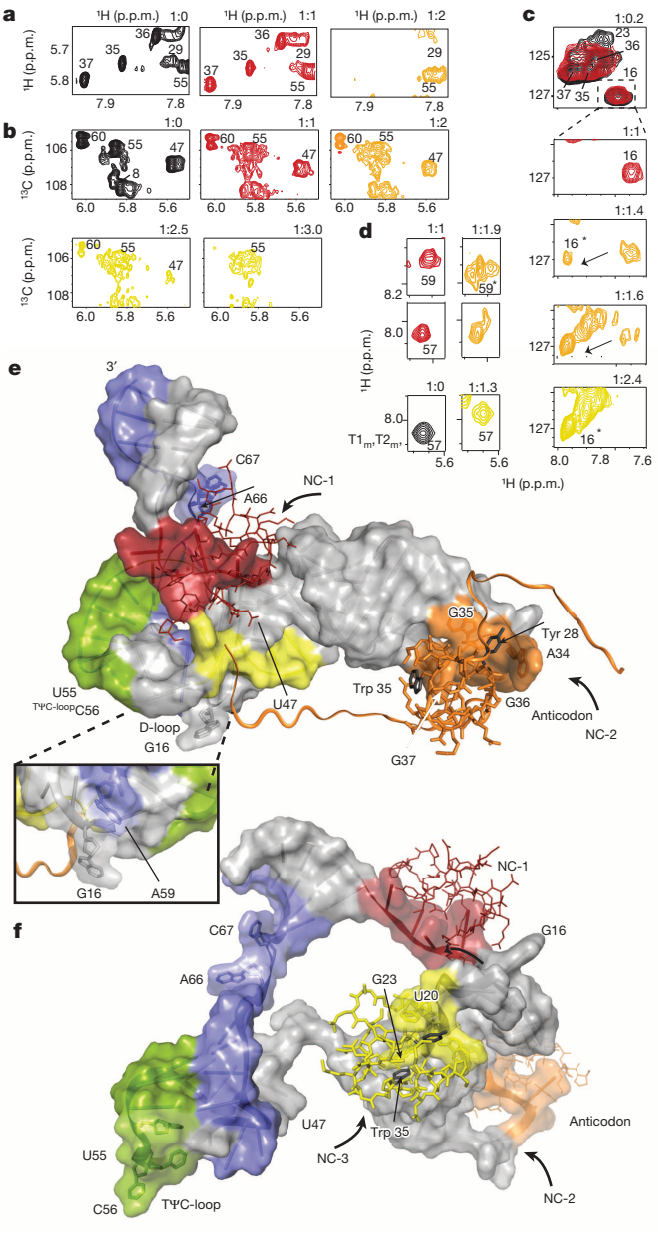


Figure 3 | Structure and NC interaction with sites T2 and T3 in $tRNA^{Pro}$. **a**, A portion of the two-dimensional nuclear Overhauser effect spectroscopy (NOESY) spectrum with H8/H1′ correlations for $tRNA^{Pro}$ showing the perturbation of anticodon residues only after titrations above 1.0 equivalent of NC, thus confirming the sequential binding mode. **b**, ^{1}H–^{13}C two-dimensional U-labelled heteronuclear multiple quantum coherence (HMQC) spectra showing that whereas the U8 resonance perturbation occurs upon the addition of one equivalent of NC, the perturbation of D-loop–TΨC signals occur only upon the third NC binding, thus indicating that the elbow contacts are maintained during the first two binding events. **c**, ^{1}H–^{13}C two-dimensional HMQC spectra showing selective perturbation of the extruded G_{16} residue in the D-loop as evidenced by a marked chemical shift change is shown by asterisks. **d**, Regions of two-dimensional NOESY for NC complexes with $tRNA^{Pro}$ and $tRNA^{Pro}$–$T1_MT2_M$. Top and middle panels show that the protected A_{57} in the TΨC loop is not perturbed, but the extruded A_{59} is affected upon the second NC binding. In the $T1_MT2_M$ mutant (bottom panel), NC binding to the third site disrupts the TΨC–D-loop interaction, resulting in a chemical shift change for residue A_{57}. Thus, the lack of A_{57} perturbation in the native $tRNA^{Pro}$ also demonstrates that the elbow region is maintained after the second NC binding. **e**, Structure of NC bound to $tRNA^{Pro}$ sites T1 and T2 via the zinc fingers. The structures show that the NC-2 protein tails can extend to contact the D-loop–TΨC-loop elbow region. Inset shows the proximity of the extruded G_{16} and A_{59} residues. For clarity, only the tails of the NC-2 protein are shown. **f**, Model of three NCs bound to $tRNA^{Pro}$ (based on the structure of NC bound to $T1_MT2_M$ $tRNA^{Pro}$; see Extended Data Fig. 9i), showing the loss of the elbow tertiary interactions upon binding of the third NC, which leads to the release of anti-PAS sequences. For clarity only the zinc-finger portion of the NC proteins are shown.

sequestered anti-PAS residues (subscript with minus sign denotes the anti-PAS position from 3′ to 5′) (Fig. 3d). NC binding to site T3 thus serves to dismantle the residual tRNA tertiary structure before primer annealing. Importantly, destabilization of the D- and TΨC-loop residues by the anticodon site T2 NC tails is maintained even after the elbow contacts are dismantled by the third NC binding event (Fig. 3c), prohibiting the freed TΨC-loop from forming an intrinsically stable structure[23] (see Extended Data Fig. 6) and ensuring that the released anti-PAS residues will remain accessible during primer annealing.

Our data demonstrate how MLV NC 'captures' specific portions of both the U5-PBS and $tRNA^{Pro}$ through high-affinity interactions with residues that are normally engaged in intramolecular stabilizing interactions and results in the subsequent 'release' of these sequestered residues, thereby reducing the energetic barrier for primer–template complex formation (Fig. 4). The combinations of liberated and pre-exposed residues within $tRNA^{Pro}$ and U5-PBS are exactly complementary and therefore poised for intermolecular base pairing. Furthermore, the complementarity of liberated sequences to regions that are already exposed in the counterpart RNA allows remodelling to occur with a limited number of NC molecules (Fig. 4). Indeed, the presence of four NC molecules is sufficient for formation of a functional U5-PBS–$tRNA^{Pro}$ complex (Extended Data Fig. 4c, d). Importantly, because the NC binding sites are perfectly positioned in close proximity to, but not overlapping with, the RNA-annealing sequences, subsequent dissociation of NC from the annealed complex is not required[24]. In fact, the presence of NC has been shown to be important for the elongation step of reverse transcription[25,26].

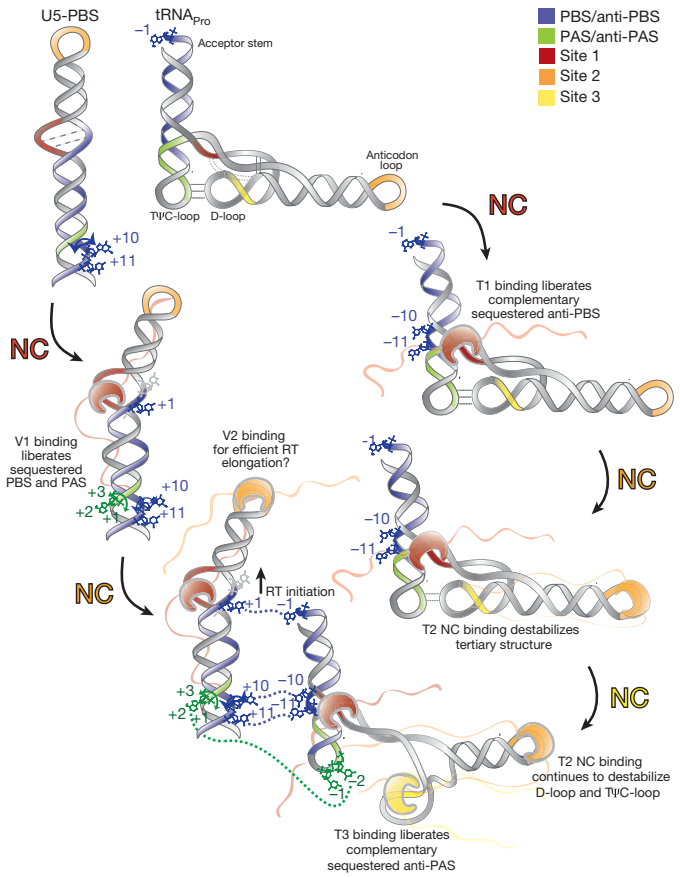


Figure 4 | Capture-and-release mechanism for NC-mediated remodelling of MLV U5-PBS and $tRNA^{Pro}$. A mechanistic model for NC zinc-finger- and tail-mediated remodelling via a step-wise, residue-specific release and residue-specific destabilization of sequestered PBS/PAS and anti-PBS/anti-PAS sequences in U5-PBS and $tRNA^{Pro}$, respectively. Upon NC binding to U5-PBS, the PBS residues $^{+1}U_{146}$, $^{+10}G_{155}$ and $^{+11}U_{156}$ are released, and the PAS residues $_{+1}G_{99}$, $_{+2}G_{100}$ and $_{+3}G_{101}$ are destabilized. Reciprocally, in $tRNA^{Pro}$, NC binding to site T1 frees anti-PBS residues $^{-10}C_{67}$ and $^{-11}A_{66}$ ($^{-1}A_{76}$ is already available in free $tRNA^{Pro}$), and binding to site T2 and site T3 renders the anti-PAS residues $_{-1}C_{56}$, $_{-2}U_{55}$ and $_{-3}U_{54}$ accessible for annealing. NC binding to the second site in U5-PBS may be important for primer extension by reverse transcriptase (RT) since stem loops with stable tetraloops have been shown to stall reverse transcription.

In addition to defining previously undiscovered roles for high-affinity NC binding events in the retroviral lifecycle, our study has implications for the location, timing and specificity of primer annealing (see Supplementary Discussion 6). Like MLV NC, some other RNA chaperones and remodellers bind with high affinity to their substrates; however, they typically require the input of additional energy for subsequent dissociation[27]. The MLV NC-mediated capture-and-release mechanism described here is distinct from mechanisms used by other known ATP-independent RNA chaperones[28,29]: During the capture-and-release RNA remodelling, NC uses high-affinity interactions to bind a limited number of sites with high specificity. Furthermore, unlike typical chaperones, which cause global destabilization to allow access to higher energy conformations, the mechanism of NC-mediated remodelling in primer annealing involves the formation of stable, lower energy complexes with RNA that cause strategic local destabilization of the regions important for annealing. Consistent with this, the thermodynamic analyses of all U5-PBS–NC and $tRNA^{Pro}$–NC interactions show high binding affinities with entropically driven profiles (see Supplementary Discussion 7). This entropy-driven, capture-and-release remodelling thus represents the first example, to our knowledge, of a new mechanism by which RNA chaperones can specifically select their specific targets from a sea of cellular RNAs.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 February; accepted 24 July 2014.
Published online 7 September 2014.

1. Harada, F., Peters, G. G. & Dahlberg, J. E. The primer tRNA for Moloney murine leukemia virus DNA synthesis. Nucleotide sequence and aminoacylation of tRNAPro. *J. Biol. Chem.* **254,** 10979–10985 (1979).
2. Wain-Hobson, S., Sonigo, P., Danos, O., Cole, S. & Alizon, M. Nucleotide sequence of the AIDS virus, LAV. *Cell* **40,** 9–17 (1985).
3. Mougel, M. *et al.* Conformational analysis of the 5′ leader and the gag initiation site of Mo-MuLV RNA and allosteric transitions induced by dimerization. *Nucleic Acids Res.* **21,** 4677–4684 (1993).
4. Paillart, J. C. *et al.* First snapshots of the HIV-1 RNA structure in infected cells and in virions. *J. Biol. Chem.* **279,** 48397–48403 (2004).
5. Wilkinson, K. A. *et al.* High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states. *PLoS Biol.* **6,** e96 (2008).
6. Levin, J. G., Mitra, M., Mascarenhas, A. & Musier-Forsyth, K. Role of HIV-1 nucleocapsid protein in HIV-1 reverse transcription. *RNA Biol.* **7,** 754–774 (2010).
7. Cordell, B., Stavnezer, E., Friedrich, R., Bishop, J. M. & Goodman, H. M. Nucleotide sequence that binds primer for DNA synthesis to the avian sarcoma virus genome. *J. Virol.* **19,** 548–558 (1976).
8. Beerens, N., Groot, F. & Berkhout, B. Initiation of HIV-1 reverse transcription is regulated by a primer activation signal. *J. Biol. Chem.* **276,** 31247–31256 (2001).
9. Beerens, N. & Berkhout, B. Switching the *in vitro* tRNA usage of HIV-1 by simultaneous adaptation of the PBS and PAS. *RNA* **8,** 357–369 (2002).
10. Thomas, J. A. & Gorelick, R. J. Nucleocapsid protein function in early infection processes. *Virus Res.* **134,** 39–63 (2008).
11. Rein, A. Nucleic acid chaperone activity of retroviral Gag proteins. *RNA Biol.* **7,** 700–705 (2010).
12. Woodson, S. A. Taming free energy landscapes with RNA chaperones. *RNA Biol.* **7,** 677–686 (2010).
13. De Rocquigny, H. *et al.* Viral RNA annealing activities of human immunodeficiency virus type 1 nucleocapsid protein require only peptide domains outside the zinc fingers. *Proc. Natl Acad. Sci. USA* **89,** 6472–6476 (1992).
14. Hargittai, M. R., Mangla, A. T., Gorelick, R. J. & Musier-Forsyth, K. HIV-1 nucleocapsid protein zinc finger structures induce $tRNA^{Lys,3}$ structural changes but are not critical for primer/template annealing. *J. Mol. Biol.* **312,** 985–997 (2001).
15. Prats, A. C. *et al.* Viral RNA annealing activities of the nucleocapsid protein of Moloney murine leukemia virus are zinc independent. *Nucleic Acids Res.* **19,** 3533–3541 (1991).
16. Rein, A., Henderson, L. E. & Levin, J. G. Nucleic-acid-chaperone activity of retroviral nucleocapsid proteins: significance for viral replication. *Trends Biochem. Sci.* **23,** 297–301 (1998).
17. Tamura, M. & Holbrook, S. R. Sequence and structural conservation in RNA ribose zippers. *J. Mol. Biol.* **320,** 455–474 (2002).

18. Nonin-Lecomte, S., Felden, B. & Dardel, F. NMR structure of the *Aquifex aeolicus* tmRNA pseudoknot PK1: new insights into the recoding event of the ribosomal trans-translation. *Nucleic Acids Res.* **34,** 1847–1853 (2006).
19. D'Souza, V. & Summers, M. F. Structural basis for packaging the dimeric genome of Moloney murine leukaemia virus. *Nature* **431,** 586–590 (2004).
20. Dey, A., York, D., Smalls-Mantey, A. & Summers, M. F. Composition and sequence-dependent binding of RNA to the nucleocapsid protein of Moloney murine leukemia virus. *Biochemistry* **44,** 3735–3744 (2005).
21. D'Souza, V. *et al.* Identification of a high affinity nucleocapsid protein binding element within the Moloney murine leukemia virus ψ-RNA packaging signal: implications for genome recognition. *J. Mol. Biol.* **314,** 217–232 (2001).
22. Theimer, C. A., Finger, L. D. & Feigon, J. YNMG tetraloop formation by a dyskeratosis congenita mutation in human telomerase RNA. *RNA* **9,** 1446–1455 (2003).
23. de Smit, M. H. *et al.* Structural variation and functional importance of a D-loop–T-loop interaction in valine-accepting tRNA-like structures of plant viral RNAs. *Nucleic Acids Res.* **30,** 4232–4240 (2002).
24. Martin-Tumasz, S., Richie, A. C., Clos, L. J. II, Brow, D. A. & Butcher, S. E. A novel occluded RNA recognition motif in Prp24 unwinds the U6 RNA internal stem loop. *Nucleic Acids Res.* **39,** 7837–7847 (2011).
25. Gonsky, J., Bacharach, E. & Goff, S. P. Identification of residues of the Moloney murine leukemia virus nucleocapsid critical for viral DNA synthesis *in vivo*. *J. Virol.* **75,** 2616–2626 (2001).
26. Liu, S., Harada, B. T., Miller, J. T., Le Grice, S. F. & Zhuang, X. Initiation complex dynamics direct the transitions between distinct phases of early HIV reverse transcription. *Nature Struct. Mol. Biol.* **17,** 1453–1460 (2010).
27. Fedorova, O., Solem, A. & Pyle, A. M. Protein-facilitated folding of group II intron ribozymes. *J. Mol. Biol.* **397,** 799–813 (2010).
28. Semrad, K. Proteins with RNA chaperone activity: a world of diverse proteins with a common task-impediment of RNA misfolding. *Biochem. Res. Int.* **2011,** 532908 (2011).
29. Dethoff, E. A., Chugh, J., Mustoe, A. M. & Al-Hashimi, H. M. Functional complexity and regulation through RNA dynamics. *Nature* **482,** 322–330 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We wish to thank the laboratory of S. Goff for the pNCS plasmid, C. Salguero for the artistic rendition of the model, as well as M. Summers and R. Gaudet for critical reading of the manuscript and the Merck Research Fellowship and Damon Runyon Cancer Research scholarship for funding.

Author Contributions V.M.D'S., S.B.M. and F.Z.Y. conceived and designed the experiments. J.A.L. and S.B.M. performed and analysed the isothermal titration calorimetry (ITC) experiments, V.M.D'S., S.B.M. and F.Z.Y. did the structural analysis, and B.W. and S.B.M. performed and analysed the virological experiments. S.B.M., F.Z.Y. and V.M.D'S. wrote the manuscript.

Author Information Coordinates and restraints for the final ensembles of the MLV U5-PBS and $tRNA^{Pro}$ structures have been deposited in the Protein Data Bank under accession numbers 2MQT, 2MQV, 2MS1 and 2MS0. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to V.M.D'S. (dsouza@mcb.harvard.edu).

A-R-T-I-S-T/GETTY

DIVERSITY

Structural approach

The field of materials science is working to broaden the range of people it attracts.

BY LEIGH KRIETSCH BOERNER

A report on diversity says that by far the most PhDs in materials science in the United States go to white men. This group also receives more and better mentoring than do female doctoral students or those from other ethnic groups. The report recommends that universities and federal agencies work to boost enrolment of graduate students from minority groups and improve their mentoring, and initiatives and programmes have already popped up around the country that aim to equalize the playing field in the discipline.

The report, *Ethnic Diversity in Materials Science and Engineering*, released in September by the US National Science Foundation (NSF), is an outgrowth of a symposium in 2012 on the state of diversity in the field. Sponsored by the NSF and the US Department of Energy, among others, the symposium found that gender and racial bias continue to affect education and workplace practices. The report was released almost two years after the workshop because it included contributions that took time to collect and organize, says Justin Schwartz, who organized and chaired the 2012 symposium and is head of the materials-science and engineering department at North Carolina State University in Raleigh.

Yet efforts are under way to mitigate bias and its effects in the field. Olivia Graeve, a materials scientist at the University of California, San Diego (UCSD), launched a small-scale research-education venture in 2012 for Latino high-school students from San Diego and Mexican students from Tijuana. The programme, called REACH, aims to motivate these students to attend university, and is run between their third and final years of high school. Participants live in the UCSD dorms for seven weeks while doing research in Graeve's and other faculty members' labs.

The first class drew five students. The second attracted 20, and next summer she expects 40, even though the programme is advertised only on its website (go.nature.com/bv3qzp). "For many of the students, this summer of research had an impact in choice of major, and also in choice of university," she says, especially for Mexican students who hope to come to the United States. Funding for her venture comes from the NSF, the UCSD chancellor's office and from her own start-up package.

Graeve says that she was fortunate to have received a great deal of support as an undergraduate, both in mentoring and in research opportunities, and was lucky to have a great mentor during her PhD at the University of California, Davis.

She has always felt that she and her research were sought after and respected wherever she has been studying or working, including San Jose State University in California; the ▶

University of Nevada, Reno; Alfred University in New York; and the UCSD. She says that she had never personally encountered overt discrimination or felt disadvantaged in her field even though she is a woman and a Latin American. "I have always been very optimistic, and I tend to have blinders on about this type of thing. I don't notice subtle things, and discrimination tends to be subtle," she says.

Nonetheless, Graeve knew that bias existed, and came face to face with it once. "At one point in my career, somebody said I got hired because I was a Latina, and I said, 'So what?'" So she started to think about ways to combat such opinions and had come up with the idea for her initiative before she attended the symposium in 2012. The symposium helped her to develop concrete ideas for implementing it, however.

UNIVERSITY OF TEXAS AT DALLAS

Magaly Spector (second from left) was involved in setting up a programme for young women in science.

STRONG APPROACH

More formally, the Partnership for Research and Education in Materials (PREM) programme is trying to boost diversity in materials-science research and education and close the mentoring gap. Launched in 2004 by the NSF and the worldwide Materials Research Society, PREM matches undergraduate and high-school students with graduate programmes throughout the United States for a variety of activities, including summer research projects, poster sessions, lectures and networking opportunities.

Yves Chabal and Magaly Spector of the University of Texas at Dallas have brought undergraduates who have gone through PREM into the university's materials-science graduate programme. They launched an outreach initiative this year with Xavier University of Louisiana in New Orleans, a historically African American university that does not have graduate programmes in the sciences, to bring its students to Dallas. As a result, the initiative has already drawn two more graduate students.

Spector, assistant to the president of strategic initiatives and a professor in practice at the University of Texas at Dallas, is the chair of the Materials Research Society diversity subcommittee, for which she sets up mentoring programmes that bring students together with graduate institutions. She and Chabal, who heads the university's materials-science and engineering department, have also created the Young Women in Science and Engineering Investigators Program, another initiative that resulted from the diversity workshop. The venture targets girls from the Dallas area, where many students struggle to find the funds to attend university, who are in their third or fourth year of high school. "These are very motivated students," Chabal says, "with no exposure to research. Some of them haven't even thought of going to college."

"If you're at an institution that does not even believe that bias plays a role, it's a giant step to develop awareness."

Teams of girls choose research projects and are paired up with graduate students and postdocs. They then work on these projects over the academic year and finish with a poster presentation. The three teams who perform best receive partial scholarships to the university. This is the programme's third year, although so far none of the winners have gone on to the university — they were seniors at the time, and had already made college plans. This year, 51 girls, mostly in their third year, are participating; of these, 9 will win scholarships.

The NSF report incorporates results from a survey conducted at the symposium that aimed to find out what is blocking students from minority groups from starting careers in materials science. The results reveal a striking difference in the type and amount of mentoring that students from minority groups receive compared with white men. For example, students from minority groups are less likely to be introduced to other researchers in the field by their advisers at conferences; they also receive less encouragement from advisers, faculty members, lab heads or department chairs, for example, to submit their research — in short, the sorts of practices that help to smooth a student's path into professional development, which can help to lead into a successful academic or industrial career.

The report found that in 2011, white people, and in particular white men, overwhelmingly earned the highest proportion of materials-science PhDs in the field, at 60% and 76%, respectively. Asians, the most successful minority group, earned around 14% of the PhDs. Degrees were rarely awarded to African American or Hispanic students, at around 3% each. Women get about 24% of the PhDs.

Schwartz helped to organize the symposium as a way to identify diversity issues at the graduate level. "There really hasn't been that much progress made at the higher levels [of education]," he says. "I thought there was a need to come at it from a different perspective."

Although the report incorporates input from materials-science researchers and students from minority groups, it does not provide suggestions for responding to or countering any bias or discrimination that they might encounter during training or in the workplace.

SELF-AWARENESS

Materials scientists at all levels need to try to become more aware of their biases, especially unconscious ones, the report recommends. "This is easy to say, not easy to do," says Eve Fine, a researcher and workshop developer at the Women in Science and Engineering Leadership Institute at the University of Wisconsin–Madison. That awareness is the first step to overcoming biases, she says. A second step is to replace the bias with a counterpoint: for example, when someone hears someone say that women are not good at maths, that person should think of a woman who is very good at maths.

An online quiz (https://implicit.harvard.edu/implicit) can help people to learn to recognize their unconscious biases, and that is the first step to dissolving them. Even considering gender and race when hanging pictures on an office wall or posting on a department web page can help to combat implicit bias, Fine says.

Although these activities can boost awareness of biases, the road to equality in the sciences is a long one, Fine concedes. "If you're at an institution that does not even believe that bias plays a role, it's a giant step to develop awareness," she says. But she believes that incremental improvement is taking place. "It's not going to be two years and the problem goes away. But I think we are seeing progress." ■

Leigh Krietsch Boerner *is a freelance writer in Bloomington, Indiana.*

ICE AND WHITE ROSES

A distant memory.

BY REBECCA BIRCH

"I always preferred white roses," Maria said, sliding an arthritis-clawed finger over a velvety red petal. The diamond in her wedding ring caught a sunbeam, casting a fractured rainbow onto the smooth face of the young woman sitting in the chair facing her. The woman's expression was guarded, her hands clenched in her lap. She perched on the front edge of the chintz-covered chair as if she were preparing to bolt.

The rainbow slid across her turned-down lips.

Maria sighed and set the bouquet of red roses on her side table next to the digitally timed pillbox. "It's kind of you to bring them."

The young woman leaned back a bit. "Mum said that you liked roses."

"Oh?" Maria asked, brightening. "Who is your mother? Do I know her?"

A frown furrowed the young woman's brow. "My mother was Sigrid Jonsson. Your daughter."

Maria blinked. Where was Sigrid? She'd last seen her … digging worms in the back garden. It had been a long time since Maria had checked on her and the child didn't always remember to stay out of the green beans and lettuce. "Will you excuse me?" she asked, reaching for her cane. "I need to be sure Sigrid's not uprooting the vegetables."

The young woman touched the back of Maria's hand. "Wait," she said. "I just came in. Everything was fine outside. Nana, please." She swallowed and her fingers pressed just a bit harder. "We need to talk."

A sharp constriction pulled at Maria's chest. Her eyes flashed to the cryopreserver on the mantel, displaying her wedding bouquet of white roses. "Is it about Harry? Has there been word?"

She never should have agreed to let him go. *Just a brief exploratory mission,* he'd said. *I'll be back before you know it.* The white roses had barely begun to crystallize. She hadn't even known she was pregnant when he left.

"No, Nana," the woman said. "There's been no word. It's about what we talked about the last time I was here."

She'd been here before? Maria squinted, looking for any familiarity in the soft face framed by pale curls. "I'm sorry, dear, but I seem to have forgotten your name?"

The woman squeezed the bridge of her nose and blinked away a bright sheen in her eyes. "I'm Margaret, Nana, and I need you to try to focus."

Focus. That was how the government

JACEY

scientists had tried to explain it. A never-before-observed phenomenon in the vast nothingness between the orbits of Jupiter and Saturn, funnelling the faint rays of the Sun into a vortex. Not a black hole. Inexplicable. Harry's ship, the *Kennedy*, had reported launching exploratory probes, then had never been heard from again.

But he'd be back. He'd sworn it by the ice-crusted roses waiting to be thawed and spread over their graves when they both had passed beyond. His homing beacon, he'd called it. His lodestone.

"Do you remember what we talked about?"

Maria blinked away the blur of tears that had formed in her eyes and tried to concentrate. This nice young woman had brought her flowers. She owed her the courtesy of trying. The light touch on the back of her hand felt familiar. Conjured a wisp of memory. The same low voice. The words …

"You want me to come with you," Maria breathed. "Somewhere far from here."

"Yes. Good. That's right."

Maria opened her eyes. "I'm not leaving."

"I know you feel comfortable here, Nana," the young woman said, tucking a curl behind one ear with her free hand, "but I've been offered a position at the Mars habitat ring. It's what I've wanted all my life. I'll be involved in research I couldn't even consider here on Earth."

Maria shook her head, chewing on her lower lip.

"There are good facilities there. They'll be able to take excellent care of you —"

"No," Maria said. "Sigrid is happy here. And when Harry returns, this is where he'll look for me."

"Mars is closer to the anomaly," the young woman said, leaning close. Her breath smelled of cinnamon, the sharp scent overwhelming the soft odour of the roses. "If that's where grandfa — Harry — is, you'll be closer to him."

"I've told you, I'm not leaving. Now please, Sigrid will be in soon and she'll be wanting her lunch." Maria turned her body pointedly away.

The woman sighed, then gathered up her handbag and rose to her feet. "I wish you'd come of your own volition, Nana, but you need to know that the courts have named me your legal guardian. You'll be coming with me …"

Maria gazed at the decorative cryopreserver, and the woman's voice faded into a low drone at the edge of her awareness. From the surface of the curved glass, she saw Harry's eyes gazing out at her, as they had since the moment the *Kennedy* lost contact. They looked different from what she remembered. When had those wrinkles appeared around the edges? Why did he look so sad?

The door closed behind the young woman with a muted thud.

Groaning, Maria struggled to her feet, leaning heavily on her cane, and crossed to the mantel. She touched the curved glass, just above her husband's brow. "Hurry home, my love," she whispered. "Sigrid and I will be waiting."

Moisture glimmered in the corner of his eye. His lashes touched the skin of her palm, whisper-soft.

The pillbox chimed and Maria turned away. The room smelled of roses and cinnamon. She swallowed her pills, then touched the bouquet beside the box, wondering where it had come from.

"I always preferred white roses," she said. ■

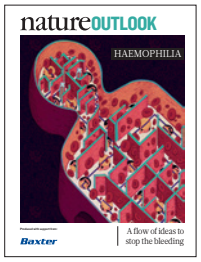
Rebecca Birch *is a Seattle resident who has been published in markets including* Nature, Grantville Gazette's Universe Annex *and* Penumbra eMag. *You can find her online at www.wordsofbirch.com.*

NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm

nature OUTLOOK

HAEMOPHILIA

27 November 2014 / Vol 515 / Issue No 7528



Cover art: Jessica Fortner

In any complex machine, the lack of a single part can lead to big trouble. That is the problem faced by the 170,000 people globally who have the bleeding disorder known as haemophilia. A genetic mutation (usually inherited) suppresses the production of proteins that make blood coagulate (see page S158). Internal bleeding into the joints causes bone degradation and excruciating pain (S170), and even mild injuries can be life-threatening.

The standard therapy is frequent infusions with blood-clotting promoters. These treatments are uncomfortable and expensive, so it is welcome news that several longer-lasting clotting factors have been developed (S162). Many people develop an immune resistance to these infused factors, but relief may be on the way in the form of anti-inhibitory pills made from plants (S166). Development of these pills depends on colonies of haemophilic dogs that serve as cooperative test subjects (S172).

Clotting-factor infusions treat symptoms of haemophilia, but gene therapy could provide a cure (S160). Research is also moving ahead on an alternative treatment strategy to remove or disable the body's anticoagulants (S168) rather than adding clotting factors.

The haemophilia community is still haunted by the traumas of blood supplies that were contaminated with HIV and hepatitis C. These experiences have led to reluctance to accept the good news that may soon be on offer, says medical historian Stephen Pemberton (S165).

To maximize its impact, this Outlook is being published in both *Nature* and *Scientific American*.

We are pleased to acknowledge the financial support of Baxter Healthcare Corporation in producing this Outlook. As always, *Nature* has sole responsibility for all editorial content.

Herb Brody
Supplements Editor

CONTENTS

COLLECTION

Nature Outlooks are sponsored supplements that aim to stimulate interest and debate around a subject of interest to the sponsor, while satisfying the editorial values of *Nature* and our readers' expectations. The boundaries of sponsor involvement are clearly delineated in the Nature Outlook Editorial guidelines available at go.nature.com/e4dwzw.

CITING THE OUTLOOK
Cite as a supplement to *Nature*, for example, Nature **Vol. XXX**, No. XXXX Suppl., Sxx–Sxx (2014).

VISIT THE OUTLOOK ONLINE
The *Nature Outlook Haemophilia* supplement can be found at http://www.nature.com/nature/outlook/haemophilia
The website features all newly commissioned content as well as a selection of relevant previously published material.

All featured articles will be freely available for 6 months.

SUBSCRIPTIONS AND CUSTOMER SERVICES
For UK/Europe (excluding Japan): Nature Publishing Group, Subscriptions, Brunel Road, Basingstoke, Hants, RG21 6XS, UK. Tel: +44 (0) 1256 329242. Subscriptions and customer services for Americas — including Canada, Latin America and the Caribbean: Nature Publishing Group, 75 Varick St, 9th floor, New York, NY 10013-1917, USA. Tel: +1 866 363 7860 (US/Canada) or +1 212 726 9223 (outside US/Canada). Japan/China/Korea: Nature Publishing Group — Asia-Pacific, Chiyoda Building 5-6th Floor, 2-37 Ichigaya Tamachi, Shinjuku-ku, Tokyo, 162-0843, Japan. Tel: +81 3 3267 8751.

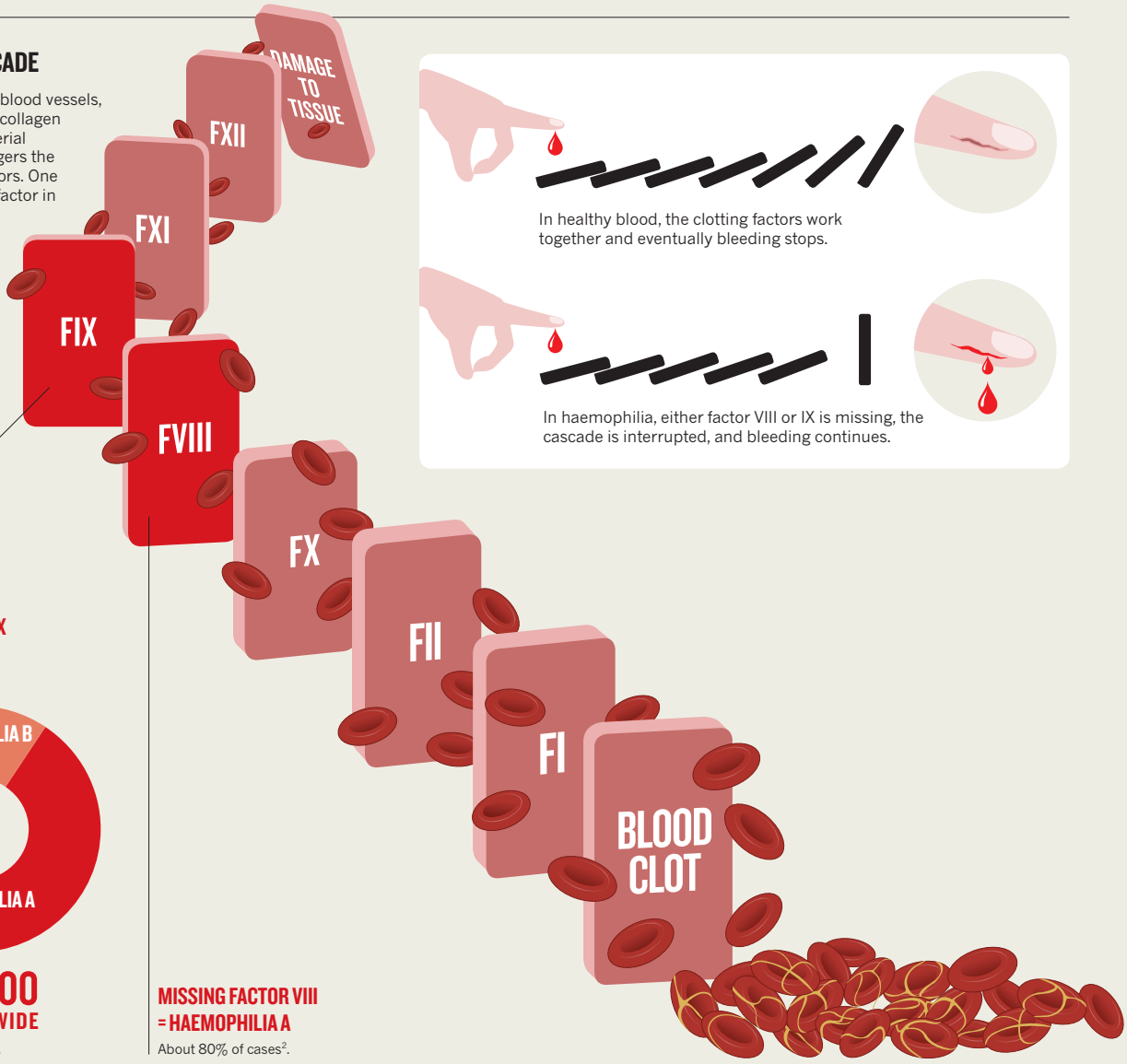
CUSTOMER SERVICES
Feedback@nature.com

BORN IN THE BLOOD

People with the inherited bleeding disorder haemophilia lack factors that cause the blood to clot. The disease affects thousands of people around the world and has even played a part in historic events. By **Neil Savage**.

COAGULATION CASCADE

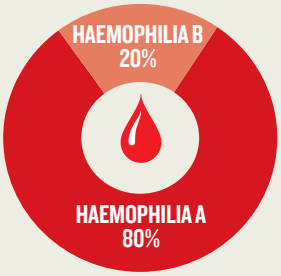
When damage occurs to blood vessels, exposure of the blood to collagen in the cell walls and material released by the cells triggers the activation of clotting factors. One factor activates the next factor in a series of events (some not depicted here) that eventually produces fibrin. Fibrin forms a mesh to hold together a plug of platelets to form a clot (platelets are a type of cell that circulates in the blood to help coagulation)[1].



In healthy blood, the clotting factors work together and eventually bleeding stops.

In haemophilia, either factor VIII or IX is missing, the cascade is interrupted, and bleeding continues.

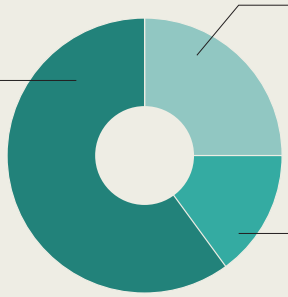
MISSING FACTOR IX = HAEMOPHILIA B
About 20% of cases[2].



AT LEAST 172,000 PEOPLE WORLDWIDE HAVE HAEMOPHILIA

MISSING FACTOR VIII = HAEMOPHILIA A
About 80% of cases[2].

OF ALL THE PEOPLE WITH HAEMOPHILIA...



60%
Have severe haemophilia (clotting factor in the blood is less than 1% of the level in normal, healthy people)[3]
Bleeding after injury, spontaneous bleeding, risk of joint damage. Median age at diagnosis is one month.

25%
Have mild haemophilia (clotting factor levels of 6–30%)[3]
Bleeding only after serious injury, accidents or surgery. Heavy menstrual bleeding in women. Median age at diagnosis is 36 months, though it is often not diagnosed until after a serious injury.

15%
Have moderate haemophilia (clotting factor levels of 1–5%)[3]
Bleeding after injury, some spontaneous bleeding, risk of joint damage. Median age at diagnosis is eight months.

TREATMENT TIMELINE

About two-thirds of the world's population lacks access to prophylaxis with clotting factors because the cost of treatment is too high.

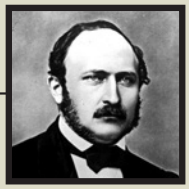
1840

1840s
First blood transfusions for bleeding episodes.

THE ROYAL DISEASE

Queen Victoria became a carrier of haemophilia through what is believed to have been a spontaneous genetic mutation[4].

Queen Victoria
1819–1901



Prince Albert
1819-1861

Victoria passed it on to one son, who died of bleeding, and two daughters.

Victoria

Edward VII

Alice
1843–1878

Alfred

Helena

Louise

Arthur

Ludvig IV

Leopold
1853–1884

Beatrice
1857–1944

Haemophilia spread throughout European royalty, including Victoria's great-grandson, Alexei Romanov, son of the last Russian tsar, Nicholas II.

The 'mad monk' Rasputin claimed to be able to treat Alexei's haemophilia, and his influence with the tsar's family is credited with contributing to the Russian revolution.

In 2009, DNA tests on Alexei's remains showed that Victoria carried haemophilia B.

Alexandra
1872–1918

Nicholas II of Russia

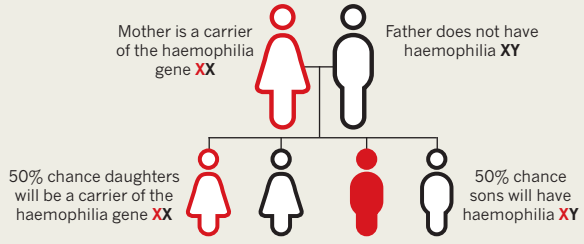
Alexei
1904–1918

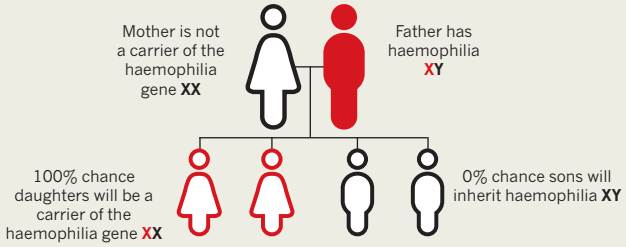
HULTON/GETTY; THE PRINT COLLECTOR/GETTY; IMAGEPAST/ALAMY; FINE ART IMAGES/HERITAGE IMAGES/GETTY; HERITAGE IMAGE PARTNERSHIP LTD/ALAMY

INHERITANCE PATTERN

The mutations causing haemophilia are carried on the X chromosome. Woman are usually carriers, with a 50% chance of having sons with haemophilia or daughters who are carriers. Men with haemophilia will have no sons who are haemophiliacs, but all their daughters will carry the gene. In rare cases, female carriers or girls with both X chromosomes affected will have haemophilia; fewer than 10% of cases occur in females. Approximately one-third of cases arise from spontaneous mutation[5].

Does not have haemophilia
Carrier of haemophilia gene
Has haemophilia





An estimated 40,000 haemophiliacs worldwide were infected with HIV, including 70–90% of those with severe haemophilia[6].

40,000 HIV

The cost of prophylaxis treatment could reach $300,000 per child per year[7].

$300,000

1900

1923 First use of plasma replacement therapy

1950s–1960 Fresh frozen plasma requires large volume transfusions

Mid-1960s Cryoprecipitate (a concentrate of blood components made from frozen plasma) allows smaller transfusions

1970 Freeze-dried clotting factors allow home-based prophylaxis

1970s–early 1980s Haemophiliacs contract hepatitis and HIV from blood products

1987 Heat-treated factor VIII reduces risk of blood-born pathogens

1989 Genetically engineered factor VIII, no risk of HIV or hepatitis

2000

1997 Genetically engineered factor IX

2014 First extended-life clotting factors approved

Future Gene therapy

References: 1. Pipe, S. W. (Ed.) The Hemophilia Report (2014) available at: www.hemophiliareport.com 2. 2012 World Hemophilia Foundation Survey, covering 91% of the world's population; 3. US National Hemophilia Foundation 4. Rogaev, E. I. *et al. Science* **326,** 817 (2009). 5. US Centers for Disease Control and Prevention 6. Starr, D. *Blood: An Epic History of Medicine and Commerce*, p346, Harper Perennial (2000). 7. Manco-Johnson, M. A. *et al. N. Engl J. Med.* **357,** 535–544 (2007).

JESSICA FORTNER

GENE THERAPY

Genie in a vector

Repairing the faulty genes that cause haemophilia could ultimately cure the disease, but it will be a tough challenge.

BY JULIE GOULD

Martin never learned to ride a bike, could not play football with his friends and wore a crash helmet when playing in the garden, just in case he bumped his head. His parents had good reason to be protective: his severe haemophilia B meant that the gentlest touch could lead to a serious, debilitating bleed. "It's very frustrating, growing up with haemophilia," says Martin. "You want to be like the other kids, but you can't."

As a result of an inherited genetic mutation, people with haemophilia B lack a protein called factor IX that is crucial for forming blood clots (see page S158). Currently, patients are treated several times a week with infusions of a concentrated version of the protein. This stops the bleeding, but it does not address the underlying cause of the disease nor does it fully remove its debilitating symptoms.

A few years ago, Martin had to stop his work as a truck driver. "I was letting the company down because I couldn't make it into work," he says. "The bleeding into my joints had made it very painful for me to move." In 2011, after 37 years of pain and joint degeneration caused by internal bleeding, Martin signed up for a clinical trial of a gene-therapy treatment at the Royal Free Hospital in London, hoping that it would provide some relief.

Rather than infusing functional clotting factors, the therapy aims to get the body to create its own. DNA with a functional factor IX gene was bundled into the molecular wrapper of a virus — known as AAV8 — then shuttled into liver cells, where factor IX is normally made.

Of the six patients who enrolled, four were able to discontinue their infusion treatments after the therapy[1]. Martin was one of them: his factor IX levels increased significantly, taking him out of the severe haemophiliac range and into the moderate group. His clotting factor levels have remained stable ever since.

The success was a crucial stepping stone for Edward Tuddenham, emeritus professor of haemophilia at University College London, who led the clinical trial. He wants to find a treatment not just for haemophilia B but for the much more common haemophilia A — but that is turning out to be a challenge.

FREEDOM OF EXPRESSION

The viral vehicle AAV8 is ideal for treating haemophilia B, but it works less well for haemophilia A. This is because the DNA encoding the clotting factor that is missing in the latter — factor VIII — is about six times larger than for factor IX, so it doesn't fit into AAV8. To make it fit, researchers often cut 4,500 base pairs out of the factor VIII gene sequence. The section they delete encodes a specific region of the protein — called the B-domain — that ensures efficient secretion of factor VIII. In its place, Tuddenham and his colleagues tried inserting a DNA sequence that is one-fiftieth of the size, but has the same function. But in a 2010 study of haemophiliac mice, these B-domain-modified treatments did not increase the level of factor VIII expressed in the blood[2]. Since then, Tuddenham has not only been trying to fix the gene but also to improve its expression.

The rate at which the factor VIII gene produces its protein is affected in part by the placement of the triplets of DNA bases — codons — that dictate where translation of the genetic material into protein should start and stop. The start and stop codons in the DNA sequence of a normal mouse or human factor VIII gene, did not promote vigorous protein production. "So we replaced them with better ones," says Tuddenham. When that was done, expression levels in a mouse model of haemophilia went from about 2% of that found in healthy mice to about 2,000%. The increase produced by the codon optimization was "enormous, truly stunning", he says.

In 2015, Tuddenham and his team hope to lead trials to test safety

NATURE.COM
Meet people living with haemophilia in our film:
go.nature.com/dzyned

and efficacy of their optimized factor VIII gene therapy for people with haemophilia A. The number of people in the trials is likely to be between 10 and 20, but even if the factor is expressed effectively in humans, there are still hurdles to overcome.

DELIVERY ON A PLATELET

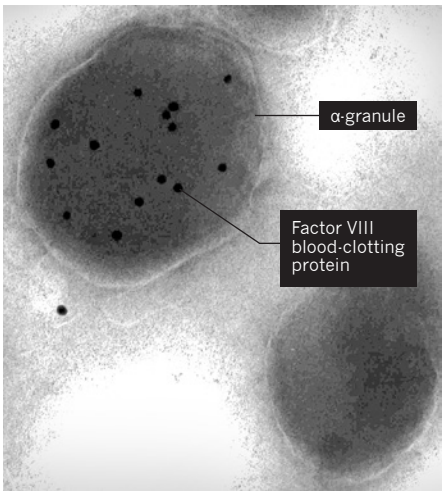
One hurdle is that AAV8 can be administered only once, because the virus triggers a strong immune response. "After one treatment with AAV8, you can't ever have a repeat dose. You are immunized against it," says Tuddenham. So although gene therapy could be a one-off cure, if the immune response is triggered before the therapy reaches the target, it is useless.

David Wilcox at the Children's Hospital of Wisconsin Research Institute in Milwaukee, hopes to get around this problem by using the body's own cells to deliver factor VIII. He is developing a way to insert functional factor VIII into structures called α-granules, which are found inside blood cells called platelets (see image, right). Platelets are the first cells to arrive at a wound site, where they rapidly begin to help form blood clots by releasing chemical messengers. Wilcox is working on modifying platelets to also release functioning factor VIII. "This removes the problem of having AAVs and factor VIII proteins floating around the rest of the body," says Wilcox, "thus avoiding any immune reactions."

"We've still got a lot to learn about gene editing in large animals before we even think about trying it in adult humans."

First, however, Wilcox has to harvest blood stem cells from the patient. He uses growth factors to coax stem cells in the bone marrow out into peripheral blood vessels, where they can easily be collected. The stem cells, which make up 2–5% of the peripheral blood sample, are then separated out in a procedure called peripheral blood stem cell apheresis and undergo gene therapy so that they contain the working factor VIII. The patient then has chemotherapy to partially suppress their existing bone-marrow stem cells before receiving a transfusion of the engineered stem cells into the blood. These cells find their way back to the bone marrow, where they will eventually produce platelets that contain functioning factor VIII.

In 2013, Wilcox tested the procedure on three dogs with severe haemophilia A, using a human factor VIII gene — and two of the dogs no longer require the usual treatment with infused factor VIII[3]. As predicted, none of the dogs showed signs of developing antibodies to the human factor VIII proteins — when the dogs received a cut, blood clots formed faster than they had without the gene therapy. "We think that the factor VIII is secreted from the platelets so quickly at the trauma site that the immune system does not have time to react before the factor VIII can start repairing the vascular injury," says Wilcox. Like Tuddenham, Wilcox's team hopes to start clinical trials inext year.



DAVID A. WILCOX & PAQUITA NURDEN

Researchers are modifying platelets to release factor VIII from α-granules at the site of injury.

But even if platelets can offer an alternative delivery vehicle, it could be an unpleasant one for patients. "I think they have a viable approach for patients with antibodies to AAVs or those affected by HIV and hepatitis," says Tuddenham, "but the doses of chemotherapy treatment before the stem-cell transplant aren't a walk in the park".

CORRECTING IN PLACE

So far, gene-therapy trials have focused on adults with the disease, but haemophilia is an inherited disease, affecting a person from birth. Unfortunately, the technique is not a viable option for children. If a child's liver were to be infused with factor VIII genes introduced through AAV, there would be an initial increase in the levels of clotting factor in the blood, as with the adults in Tuddenham's 2011 trial. But as the child grows, the expression levels would decrease when new liver cells are produced without the functioning factor VIII gene are produced, says haematologist Katherine High, at the Children's Hospital of Philadelphia in Pennsylvania.

In theory, Wilcox's method might work in children because the functional clotting-factor genes have been integrated into the stem cell's genome and will be passed on to daughter cells. In practice, however, no responsible physician would expose an infant or child with a non-lethal disease such as haemophilia to chemotherapy.

A promising way to avoid these problems is *in vivo* genome editing, in which mutant genes are corrected *in situ* rather than replaced. This could potentially work at any age — but the earlier in life such a treatment is available, the better, as the benefit would be lifelong.

Conceptually, this approach is as simple as setting up a biological tool to cut out the mutated area of the genome, then another to insert a corrected template, says Merlin Crossley of the University of New South Wales, Sydney, Australia. Crossley sees gene-editing therapies as the best potential tool for curing haemophilia.

This could be particularly beneficial for children: as the liver grows, the new daughter cells would contain the functioning clotting-factor gene. The clotting factor would then be recognized as part of the body, and could ultimately eliminate the child's haemophilia. "The replacement template is cloned from healthy patients and wouldn't be attacked by the immune system because it isn't considered as foreign," Crossley says.

A 2011 study in mice[4] by High provided strong evidence that genome editing is a viable option. Immediately after birth, one set of mice was given Tuddenham's style of gene-transfer therapy; a second set was given the genome-editing treatment. High discovered that the levels of functioning clotting proteins in mice receiving the genome-editing treatment stayed high even after a portion of their liver was surgically removed; in the mice receiving gene therapy, by contrast, factor levels decreased. "This is the advantage of this treatment, especially for children," says High.

Genome editing has to be precisely targeted to the mutation to be repaired, and the sheer number of mutations for haemophilia A — more than 2,000 — makes this a challenge.

Both High and Tuddenham believe that in the short term, genome editing is not the answer. "The gap between proof-of-principle experiments in mice to clinical trials in humans for gene-transfer therapy was 14 years," says High. "And we've still got a lot to learn about gene editing in large animals before we even think about trying it in adult humans, let alone infants."

Having his haemophilia reduced to a moderate level has improved Martin's quality of life tremendously. He has needed the standard infusion treatment fewer than ten times since the gene therapy, and says that "only one of those occasions was a serious bleed". He says that signing up for the trial was not an easy decision, because there were not any other similar trials on which to base his decision. But he believes that his successful experience should help to encourage people to participte in future studies. "You go from a position of knowing what you are, how you are and how to deal with it, to a position of complete uncertainty," he says. "So I hope that the uncertainty is reduced for other patients when they hear about our experiences." ■

Julie Gould *is the editor of Naturejobs.*

1. Nathwani, A. C. *et al. N. Engl. J. Med.* **365,** 2357–65 (2011).
2. McIntosh J. *et al. Blood.* **117,** 798–807 (2011).
3. Du, L. M. *et al. Nature Comm.* 4, 2773 (2013).
4. Li, H. *et al. Nature* **475,** 217–21 (2011).

WORLD FEDERATION OF HEMOPHILIA

Boys with haemophilia receive a blood-clotting factor by intravenous injection (also referred to as an infusion).

CLOTTING FACTORS

Stretching time

Extending the life of clotting factors may improve quality of life for people with haemophilia.

BY NEIL SAVAGE

For the parents of a child born with haemophilia, the diagnosis comes with both good and bad news. The good news is that the child, at least if he (or rarely, she) is born in the developed world, can expect a near-normal lifespan, up from a mere 20 years in 1970. The bad is that the parents must teach themselves to find their child's veins, insert a needle and infuse him with a clotting factor to replace what he lacks. Parents must infuse a toddler as often as every other day, and children with haemophilia will have to continue that treatment for the rest of their lives.

But treatment is getting easier. Down the road, gene therapy and other approaches look likely to bring longer-term treatments for patients with the rare bleeding disorder. For now, improvement in treatment lies in the emergence of new, longer-lasting replacements for the blood-clotting factors missing from the blood of people with the condition. These therapies could stretch the time between infusions to days or even weeks. The first two such treatments were approved by the US Food and Drug Administration (FDA) earlier this year, and more are in the pipeline, with some expected to be approved in 2015. As these therapies emerge, dealing with haemophilia will become less troublesome (see 'Drugs to help the blood'). This could increase compliance with treatment, reduce complications — and perhaps even allow some people to live almost as if they were free of the disease.

Replacing the clotting ability lacking in haemophilia has been the treatment since the 1840s, when attempts were made to treat people with the disease by transfusion with whole blood from people with normal clotting. By the end of the 1960s, freeze-dried concentrates of clotting factors were available for home use, to prevent spontaneous bleeding. In the 1990s, treatment leapt forward again, with donated plasma being replaced by clotting factors manufactured through recombinant DNA-technology, eliminating the transmission of viral diseases that had devastated the haemophiliac community in the 1970s and 1980s.

But prophylactic treatment still has its problems. The clotting factors do not last very long in the body. Depending on the person, the amount of factor VIII — the protein missing in haemophilia A — in the bloodstream drops by half in a mere 8–12 hours. Factor IX — which people with haemophilia B lack — lasts longer, 18–24 hours. Those short half-lives mean that most people with haemophilia must transfuse themselves every two or three days. And inserting a needle directly into a vein can be difficult. "Adherence to therapy is not great, because you have to inject yourself, and it's a hassle," says David Lillicrap, a professor of pathology and molecular medicine at Queen's University in Kingston, Ontario, Canada.

One 2001 study suggested that up to 40% of

people with severe haemophilia do not follow the prophylactic schedule[1]. Those people are more likely to develop spontaneous bleeding that causes joints to fill with blood and results in progressive damage similar to arthritis. They can also develop intracranial bleeding, which can cause brain damage and even death.

Drug companies have responded with clotting factors that last longer, making the time between infusions greater. Biogen Idec, based in Cambridge, Massachusetts, has two such factors approved by the FDA this year. Eloctate, for haemophilia A, was approved in June and is recommended for an initial infusion once every four days, with a physician adjusting that up to five days or down to three as appropriate. Alprolix, the company's treatment for haemophilia B approved in March, promises infusions once a week, and perhaps every ten days or two weeks in some patients. Other versions of the clotting factors from other drug developers are showing similar extensions of lifetimes.

"It's a big improvement," says Timothy Nichols, a cardiologist and pathologist who studies haemophilia at the University of North Carolina at Chapel Hill. "It's not *no* treatment, but it is a lot easier than sticking a needle in your kid three times a week."

Steven Pipe, a paediatric haematologist at the University of Michigan's C. S. Mott Children's Hospital in Ann Arbor, agrees that the progress is significant. In particular, work that is stretching the lifetime of factor IX by three to five times is "really transformative", he says. And because half-lives can vary between patients, "at high doses, you could probably in some individual cases get a month's worth of factor IX," Pipe says.

BORROWED TIME

The trick to extending the half-lives of clotting factors is to interfere with the body's natural mechanisms for flushing them away. There are three very similar approaches, each of which extends half-life by about the same amount for the respective clotting factors. The only real difference is between factor IX, for which the techniques are offering extensions long enough to make a substantial difference in treatment, and factor VIII, for which the improvement has been more modest. Unfortunately, haemophilia A, which is caused by factor VIII deficiency, is about four times as common as haemophilia B.

Two of the techniques piggyback on the half-lives of other longer-lived proteins that occur naturally in the body. One such is immunoglobulin, a large Y-shaped protein with a half-life of about three weeks. The stem of the Y is known as the Fc region. When a clotting protein is fused to an Fc region, the body treats the clotting factor more like an immunoglobulin, and allows it to stick around for longer, although not for as long as a complete immunoglobulin molecule.

DRUGS TO HELP THE BLOOD

A number of treatments to aid blood clotting are in clinical trials or have been approved this year.

	Product	Approach	Company	Half-life (hours)	Status
Factor VIII infusions (for haemophilia A) Conventional infusion half-life: 8-12 hours	Eloctate	Fc fusion protein	Biogen Idec	20	FDA approved in June 2014
	BAX 855	PEGylation	Baxter International	19	Submission for approval planned for late 2014
	BAY94-9027	PEGylation	Bayer	19	Submission for approval planned for mid-2015
	N8-GP	PEGylation	Novo Nordisk	19	Submission for approval planned for 2018
Factor IX infusions (for haemophilia B) Conventional infusion half-life: 18–24 hours	rIX-FP	Albumin fusion	CSL Behring	92	In clinical trials
	N9-GP	PEGylation	Novo Nordisk	110	Submission for approval planned for 2015
	Alprolix	Fc fusion protein	Biogen Idec	87	FDA approved in March 2014

FDA, US Food and Drug Administration.

For factor VIII, Fc fusion extends the half-life from a maximum of about 12 hours to about 18 hours. Factor IX, which has a longer half-life to begin with, shows a more dramatic increase, from one day to five days.

Both the approved Biogen drugs are based on Fc fusion. Similar fusion drugs have been on the market to treat other diseases for many years, for example the rheumatoid arthritis drug Etanercept, which was approved by the FDA in 1998. Jerry Powell, the retired director of the Hemophilia Treatment Center at the University of California, Davis, says that the success of those drugs suggests that this is a safe approach to altering the clotting factors.

BIOGEN IDEC

Coagulation factor IX, used to treat haemophilia B

A similar approach, which is being pursued by CSL Behring, based in King of Prussia, Pennsylvania, is to fuse the clotting factors with albumin. Albumin is a major protein of blood plasma and, like immunoglobulin, has a half-life of about 20 days. Phase I safety studies of factor IX fused to albumin showed a fivefold increase in half-life, up to about four days. Unfortunately, attempts to do the same with factor VIII have been unsuccessful. Powell says that the albumin seems somehow to interfere with the normal activity of that clotting factor.

"These are really big molecules," he says. The activity of factor VIII in action, he adds, is so complex that it resembles a dancing elephant — too easily thrown off its rhythm when something else is attached. "If you put the wrong kind of contraption on the elephant, it doesn't dance as well."

The third strategy takes a slightly different approach. Instead of marrying the clotting proteins to a natural substance in the body, they are attached to synthetic polyethylene glycol (PEG) molecules (see 'PEGylation protection'). The PEG forms a sort of 'watery cloud' around the protein, protecting it from various mechanisms that would break it down; for instance, PEG prevents the clotting factors from binding to protein-specific receptors that would normally clear them away. PEG is eventually flushed from the body through the kidneys and liver, but before then it gives the clotting factors a new lease of life. Three large drug companies — Bayer in Leverkusen, Germany, Baxter International in Deerfield, Illinois, and the Danish company Novo Nordisk in Bagsvaerd — have all developed PEGylated factor VIII with a half-life of roughly 19 hours. Baxter expects to submit its product for regulatory approval by the end of this year, Bayer next year, and Novo Nordisk by 2018.

Novo Nordisk is also testing a PEGylated factor IX that has shown a half-life of 110 hours in clinical studies. The company says that it hopes to submit that drug for approval next year.

Up to now, tests have not shown much difference, in safety or effectiveness, between the three approaches. There are concerns that PEG might accumulate in the liver or kidneys over years of use, but studies of PEG have found it to have very low toxicity, and Powell thinks that those fears are exaggerated[2]. "PEG's been around a long time, there's a lot of toxicology and all the toxicology indicates no concern," Powell says. And if, as he expects, gene therapy replaces these treatments in the next decade, patients will in any case not have lifetime exposure to PEG.

One barrier to haemophilia therapy is the tendency of factor VIII to prompt the body into producing anti-factor VIII antibodies,

known as inhibitors. For a person with haemophilia A, factor VIII is a foreign substance, and the immune system can see it as a threat. About 30% of people with haemophilia A develop inhibitors, and once they do, treating their bleeding becomes much more difficult. Only about 4% of people with haemophilia B develop inhibitors to factor IX.

There is a lot of worry, Pipe says, that altering factor VIII to extend its half-life could make the inhibitor problem worse. "Everyone treads lightly in the factor VIII field, because there is such a fear of immunogenicity with any change of the molecule," he says. "There's no question with the current strategies that all of them have sort of hit a ceiling. If we're really going to overcome that ceiling, you are going to have to accept more dramatic changes to the molecule."

PEG may prove helpful in that regard. Studies dating back to the 1970s have shown that PEGylation can reduce the chances of a foreign protein stimulating an immune reaction, although the effect has not yet been proved in people with haemophilia. "That'd be a huge breakthrough if that were true," Powell says.

CONSTANT CASCADE

One researcher might have worked out a way to avoid the inhibitor issue almost entirely, by developing a different molecule to take the place of factor VIII in the clotting cascade.

Normally, once activated by previous steps in the cascade, factor VIII grabs hold of both factor IX and factor X, bringing them together to perform the next steps in the cascade. Midori Shima, director of the Hemophilia Center at Nara Medical University in Japan, has created a 'bispecific' antibody to do the same job.

Antibodies are immunoglobulins, and the upper arms of these Y-shaped proteins are designed to bind specifically to another molecule. Shima has created an antibody with one arm that binds to factor IX and the other to factor X, pulling the two together so that the clotting cascade can continue. The bispecific antibody has a half-life of about 30 days, much longer than the 12-hour upper limit of factor VIII, Shima says. Chugai Pharmaceuticals, based in Tokyo, and Hoffman-La Roche, based in Basel, Switzerland, are working on developing his findings into a treatment.

The researchers have not yet released the results of their phase II initial clinical trials, but Shima says that in the patients with haemophilia they looked at, bleeding frequency decreased dramatically. Among six people receiving the lowest dose of the treatment, who had each had 20–60 episodes of bleeding in the 12 weeks before the trial, two had no bleeding episodes at all during the 12 weeks of the trial. And out of 64 patients, only one developed an inhibitor. The team is planning a larger, phase III trial.

One bonus of this treatment is that because of the nature of the antibody, it does not have to be delivered intravenously, but instead can be injected under the skin, like insulin. "We think we can change the whole concept of haemophilia treatment," Shima says.

Lillicrap agrees. "That bispecific antibody would be hugely disruptive if it works," he says. "We'll know within the next couple of years whether it delivers on the promise which so far it's shown."

Treatments with extended half-lives may provide benefits beyond the convenience of less-frequent infusions and the potential increase in the number of people who stick to their treatment regime. If people under treatment now keep to their current schedule with the extended-life products instead of taking fewer infusions, the increased concentration of clotting factors in their blood could improve their quality of life even further.

"It is a lot easier than sticking a needle in your kid three times a week."

When patients have an infusion of clotting factor every 48 hours, the concentration of clotting factor initially reaches 100% of normal levels and stays there for about 12 hours. For the next 36 hours, it is high enough to be useful, but below normal. For the last 6 to 8 hours, the level is very low, less than 5%, Pipe says. Physicians try to keep the lowest level, the trough, from falling below 1% of the amount a non-haemophiliac person has circulating in their blood, enough to prevent spontaneous bleeding.

But if the trough level can be higher, it might make life easier for the patients, allowing them to, for instance, take up athletics with less fear of injury. "Ideally, you'd like to have zero bleeding," Pipe says. "What is the threshold for that I don't think anybody knows." Still, there would be substantial benefit from a less-than-perfect level of clotting factor. "If you could maintain a level of 10% or 15%, you would probably eliminate all joint disease," he says.

Lillicrap hopes that the emergence of several therapies means that it will make economic sense for drug companies to provide treatments to poorer parts of the world that have not been able to afford them. "No longer are people thinking about these therapies being only Western European and North American therapies," he says. If pharmaceutical companies are pouring money into this research, he thinks that it is at least in part because they can see a worldwide profit benefit.

For all the advantages of these extended-life molecules, the researchers predict that they will be supplanted in perhaps a decade by advances in gene therapy, which will enable people with haemophilia to produce their own clotting factors. But in the meantime, trading current therapies for longer-lasting ones can improve patients' lives. "As a bridging therapy between the really good outcomes we have currently and maybe a cure down the line," says Pipe, "I think the extended-half-life molecules are a perfect transition." ■

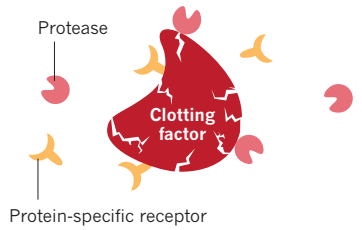
PEGylation protection

A key advance in haemophilia treatment is to prolong the effectiveness of the injected coagulation-promoting proteins (clotting factors) by shielding them from destruction.

BEFORE

Unprotected molecule

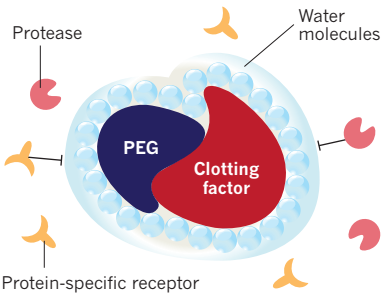
Under normal circumstances, proteases and protein-specific receptors break up the clotting factor and rapidly clear it from the bloodstream.



AFTER

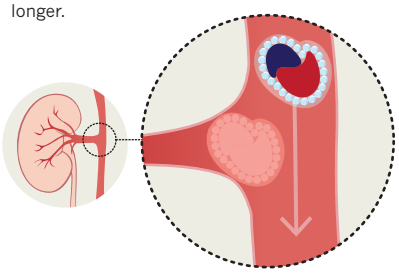
Microscopic shield

In PEGylation, molecules of polyethylene glycol (PEG) are attached to the clotting factor. The PEG molecules bring with them water molecules, which shield the clotting factor from attack.



Too big to discard

The watery cloud makes the factors too big for the kidneys' filtration mechanism, so the molecule circulates in the bloodstream for longer.



Neil Savage *is a freelance writer based in Lowell, Massachusetts.*

1. Hacker, M. R., Geraghty, S. & Manco-Johnson, M. *Haemophilia* **7,** 392–396 (2001).
2. Webster, R. *et al. Drug Metab. Dispos.* **35,** 9–16 (2007).

PERSPECTIVE


HEATHER VAN UXEM LEWIS

The fix is in

History explains why people with haemophilia, and their physicians, are cautious to believe that a cure is in sight, says **Stephen Pemberton**.

In 2011, a remarkable study[1] in the *New England Journal of Medicine* detailed the successful treatment of six adults with haemophilia B, which is caused by a deficiency in the coagulation protein known as factor IX. All of the participants were able to eliminate or reduce the frequency of clotting-factor-replacement injections — the current standard treatment for the disease — after their livers began producing functional levels of factor IX. The experimental therapy came in the form of an adeno-associated virus (AAV) carrying a gene that encodes instructions for production of normal levels of human factor IX. Three trials of AAV-mediated gene transfer in patients with haemophilia B are ongoing, with high expectations.

After more than 20 years of research on gene transfer, it is a promising time for haemophilia therapies. It now seems likely that a single-dose treatment for haemophilia B using an AAV or another gene-transfer technique will be a viable option for many people in the next decade or two.

Yet haemophilia researchers are not inclined to speak enthusiastically of a cure. Part of that caution comes from recognition that there are still problems to solve. For example, some 40% of people with haemophilia B would find no refuge in an AAV treatment because they produce antibodies that attack and neutralize this virus[2].

And even if that problem were solved, the treatment would apply only to those with haemophilia B. The more common form of the condition, haemophilia A, stems from a deficit in another protein — factor VIII — and the gene for that protein is a more difficult target. Regardless of the type of haemophilia, researchers remain hesitant about gene therapy owing to the unresolved ethical issues that arose decades ago.

RESEARCHERS ARE HESITANT ABOUT GENE THERAPY OWING TO THE UNRESOLVED ETHICAL ISSUES.

The unfettered optimism that characterized the early years of gene-therapy research came to a screeching halt in 1999, when 18-year-old Jesse Gelsinger died in a phase I clinical trial at the University of Pennsylvania in Philadelphia. Gelsinger had undergone an experimental gene transfer for his otherwise treatable metabolic disorder. His death, along with a series of other harmful events in early gene-therapy trials for a variety of diseases, threatened the whole field.

Haemophilia specialists who were engaged in gene-transfer studies were more guarded than most of that era's self-proclaimed gene doctors[3]. The source of their reserve goes beyond the cautious optimism that characterized such research after 1999; it is grounded instead in the long and troubled experience that the haemophilia community has had with technological fixes.

By the late 1970s, a therapeutic revolution had transformed haemophilia from an obscure hereditary malady into a manageable disease[4]. But the glory of this achievement was tragically short-lived. The same clotting-factor-replacement therapies that delivered a degree of normality to the lives of people with haemophilia brought unexpected and fatal results: tens of thousands of people with haemophilia were diagnosed with transfusion-related HIV/AIDS in the 1980s and with hepatitis C virus (HCV) in the 1990s.

The memory of tainted transfusions still haunts those who have, or work with, haemophilia. Add Gelsinger's death into the mix and it is clear why specialists are debating thorny ethical problems, such as when to try out AAV-mediated gene transfer on children. Gene therapy is not even the most promising treatment for haemophilia on the immediate horizon. The biotechnology industry is producing recombinant-clotting-factor products for both haemophilia A and B that can limit bleeding episodes with less-frequent injections (see page S162).

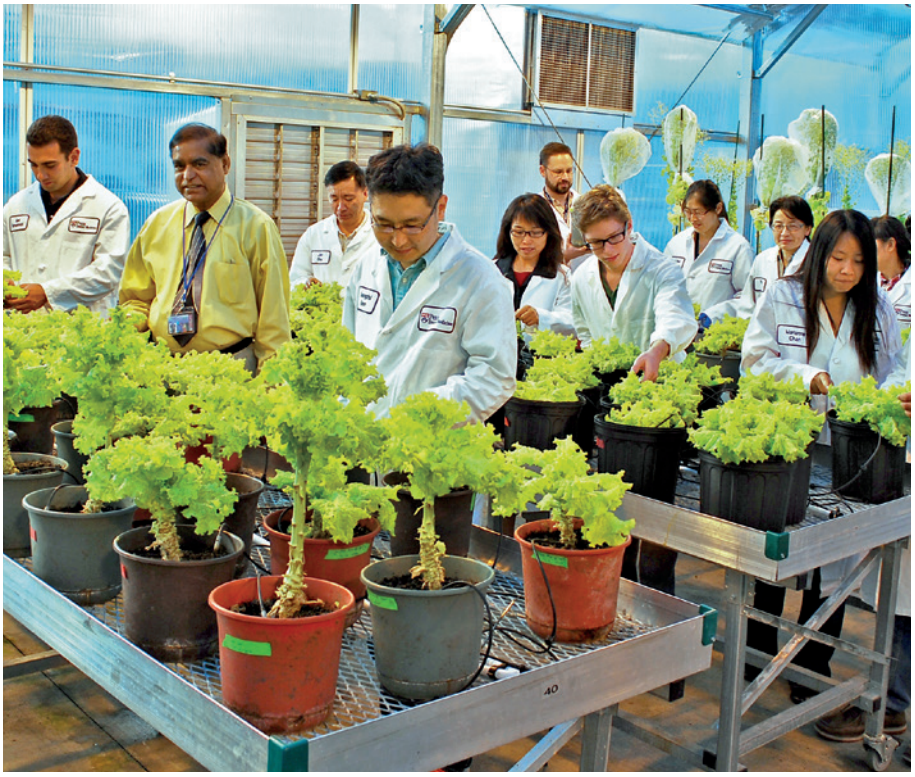
But the lure of a less-intrusive form of treatment raises a historical spectre of its own. It was this same desire for convenience that led many haemophilia physicians and patients in the United States in the 1980s to continue using clotting-factor concentrates that had a high risk of HIV contamination rather than switch back to older, more cumbersome but less risky forms of plasma-replacement therapy. Thousands of people with haemophilia contracted HIV and HCV because of this acculturated preference[4].

Finally, there is the difficulty of making costly treatments available to the vast majority of the world's haemophilia patients who live in low income countries. About 75% of people with haemophilia still receive inadequate treatment, particularly in less-developed nations where clotting-factor therapy is limited[5]. An effective gene therapy could well offer these underserved patients their first chance at effective intervention[6].

History suggests that the fix will not lie in just one solution, but will be contextual and messy. The wants and needs of people with haemophilia in the developed world might not be the same as for those in low income countries. Yet social justice demands that there be equity in access to treatment. The transfusion scandals of the past remind us of the importance of bringing together patients and treatment professionals with stakeholders from industry and public health to weigh the various technological fixes. If such discussions had taken place in the 1970s and 1980s about the known problem of transfusion-related hepatitis B, the haemophilia community would not have been blind-sided by the emergence of HIV and HCV. ■

Stephen Pemberton *is a medical historian at the New Jersey Institute of Technology in Newark, and author of* The Bleeding Disease: Hemophilia and the Unintended Consequences of Medical Progress.
e-mail: stephen.pemberton@njit.edu

1. Nathwani, A. C. *et al. N. Engl. J. Med.* **365,** 2357–2365 (2011).
2. High, K. H., Nathwani, A., Spencer, T. & Lillicrap, D. *Haemophilia* **20** (Suppl. 4), 43–49 (2014).
3. Wailoo, K. & Pemberton, S. *The Troubled Dream of Genetic Medicine: Ethnicity and Innovation in Tay-Sachs, Cystic Fibrosis, and Sickle Cell Disease* (Johns Hopkins Univ. Press, 2006).
4. Pemberton, S. *The Bleeding Disease: Hemophilia and the Unintended Consequences of Medical Progress* (Johns Hopkins Univ. Press, 2011).
5. Mannucci, P. M. *Haemophilia* **17** (Suppl. 3), 1–24 (2011).
6. High, K. A. & Skinner, M. W. *Mol. Ther.* **19,** 1749–1750 (2011).



KWANG-CHUL KWON, JIN SU & HENRY DANIELL, UNIV. PENNSYLVANIA

Blood-clotting factors produced by these lettuce plants could eliminate the problem of immune rejection.

IMMUNOLOGY

Oral solutions

Pills made from lettuce leaves could help to prevent one of the most serious complications of haemophilia treatment.

BY ELIE DOLGIN

The food in Anita's bowl is not your average dog chow. Although the dish contains pellets and wet food, there is also a sprinkling of green powder — the product of a trailblazing experiment to address a potentially lethal complication of haemophilia treatment. Anita, so named because her red coat reminded breeders of the character from the animated film *One Hundred and One Dalmatians*, is a keagle (a mix of a beagle and a Cairn terrier) with haemophilia B.

Like people with this rare genetic disorder, Anita is naturally deficient in factor IX, a protein that helps the blood to form clots. When treated with replacement coagulation proteins, the dog naturally develops antibodies, or inhibitors, against the therapy — a problem that is also seen in some 5% of humans with haemophilia B. In these people, the immune system identifies the therapeutic protein as dangerous, causing the body to stop accepting the protein as a normal part of the blood, and destroys it before it can stop the bleeding. Continuing to take factor-replacement therapies can result in life-threatening allergic reactions, such as anaphylaxis.

The problem is even worse with haemophilia A, a disease that is four times more common than haemophilia B and in which the missing link in the coagulation chain is a protein called factor VIII. Around 30% of people with haemophilia A develop antibodies against replacement factor VIII.

Therapies are available to eliminate these antibodies. Some people, for example, undergo an intensive treatment called immune tolerance induction therapy, which involves regular intravenous administration of coagulation factors. But this is time consuming and costly (around US$1 million for an average five-year-old patient), and the treatment works in only about three-quarters of patients. "The challenges of treating haemophilia with inhibitors are just staggering," says Timothy Nichols, director of the Francis Owen Blood Research Laboratory at the University of North Carolina at Chapel Hill, which maintains the colony of haemophiliac dogs to which Anita belongs (see page S172).

Inducing immune tolerance in people who have developed inhibitors is one approach. But avoiding the problem altogether would be even better. "If you can prevent antibody formation in the first place, by finding some way of producing immunological tolerance that gets around that type of protocol, that would be a major advantage," says David Lillicrap, a clinician and researcher who specializes in bleeding disorders at Queen's University in Kingston, Ontario, Canada.

The green powder in Anita's dish might do just that. The oral treatment is a concentrate of freeze-dried lettuce-leaf cells, each containing around 10,000 chloroplasts — the organelles responsible for photosynthesis — that have been genetically engineered to produce factor IX. These proteins cannot themselves be used to prevent bleeding episodes, because the cellular machinery found in plants cannot package the human clotting factors into the biologically active form. What they can do, however, is prevent the immune system from mounting an attack against subsequent therapy.

The researchers behind the bioengineered lettuce have shown that inhibitor formation and severe allergic reactions can be prevented in mice by feeding the animals with a product based on these plants[1,2]. If the strategy also works in Anita and her kennel mates — and ultimately in humans — it could form the basis of the first product to protect against the immune responses associated with haemophilia treatment.

Anita is one of only two dogs to have received the bioengineered lettuce. "So far, it's going very well," says lead researcher Henry Daniell, director of translational research at the University of Pennsylvania School of Dental Medicine in Philadelphia.

AN ACT OF TOLERANCE

In 2006, Lillicrap demonstrated that a simple oral treatment could train the immune system not to produce inhibitors. Working with a mouse model of haemophilia A, he and his colleagues gave the mice a purified fragment of the human factor VIII protein, through the nose or mouth. The researchers found that the treatment afforded some protection against antibody development after factor VIII replacement therapy[3]. But the approach did not deliver sufficient amounts of the factor to immune cells in the gut or nasal passage to fully quash inhibitor formation.

Daniell came up with an improved delivery system. He focused first on haemophilia B. Adapting a technique[4] that he had previously developed to delay the onset of type 1 diabetes, Daniell and his group genetically modified tobacco plants to express human factor IX in their chloroplasts. (Daniell has since switched to using lettuce.)

Chloroplast DNA is separate from the genome DNA in the plant nucleus, and the large numbers of these tiny organelles in the

cell allow huge volumes of the coagulation protein to accumulate in each tobacco leaf. Once ingested, the plant cell wall protects the coagulation protein from being destroyed by stomach acid. Gut microorganisms farther down the digestive tract then chew away at the cell wall, releasing the clotting-factor protein.

To target the proteins to the immune system, Daniell then attached a second protein that has high binding affinity for a receptor found on the inside of the human gut. With this fused construct tethered to the intestinal wall, the coagulation protein could be absorbed into the body and processed by the specialized cells in the immune system that induce tolerance.

Working with Roland Herzog, a molecular biologist at the University of Florida in Gainesville, Daniell then tested the plant-based product in animal models. In 2010, they showed that oral delivery of factor IX expressed in chloroplasts in this way led to almost undetectable inhibitor levels in mice, and no sign of anaphylactic shock[1]. "The mice are healthy, they show no allergic responses and they don't form the inhibitors," Herzog says. "That's pretty exciting."

Daniell then modified the tobacco leaves to express factor VIII and shipped powders of the leaves to Herzog. Earlier this year, the two researchers and their teams documented[2] suppression of inhibitor formation and even reversal of pre-existing inhibitors in mouse models of haemophilia A.

INHIBITORY CONTROL

Other strategies being pursued to prevent the formation of inhibitors of clotting-factor therapy include immunosuppressants and drugs that deplete specific immune cells. However, these therapies have many side effects, including increased susceptibility to infection.

A potentially safer option comes from Selecta Biosciences, a company in Watertown, Massachusetts. Selecta has developed a nanoparticle delivery system in which an immune-modifying compound is contained in biodegradable plastic particles just 150 nanometres across. When injected together with factor VIII into mouse models of haemophilia A, the nanoparticles deliver their payload to cells in the lymphoid tissue that are responsible for initiating immune responses. These cells, in turn, instruct factor-VIII–specific immune cells to become tolerant to the coagulation protein, resulting in suppression of misdirected antibody responses to the replacement therapy — all without affecting the rest of the immune system.

David Scott and his colleagues at the Uniformed Services University of the Health Sciences in Bethesda, Maryland, teamed up with Selecta to show that inhibitors remained undetectable for at least six months after treatment with the nanoparticle formulation[5]. "This underscores the point that we're actually teaching the immune system to become tolerant to factor VIII," says Selecta's chief scientific officer, Takashi Kei Kishimoto.

The nanotechnology approach that is being tested for inhibitor control could also improve the haemophilia treatment that is now at the cutting edge of clinical research: gene therapy. Using the standard gene-therapy approach, researchers have shown that they can achieve long-term expression of factor IX in adults with haemophilia B at sufficiently high levels to convert the bleeding disorder into a mild disease (see page S160). There has so far been no reported evidence of inhibitor formation in the small number of human participants in clinical trials for this viral therapy[6].

KWANG-CHUL KWON, JIN SU & HENRY DANIELL, UNIV. PENNSYLVANIA

Green power: from leaf to powder to capsule.

Still, the standard form of liver-targeted gene therapy carries a range of potential complications, including the risk of harmful mutations and of the body mounting an immune response against the viral vectors used to carry the correct forms of the defective genes responsible for haemophilia. That is why several research groups are attempting to replace viral vectors with nanoparticles that can deliver gene therapies as 'DNA pills'.

PILL PROTECTION

DNA pills combine DNA plasmids — circular pieces of bacterial DNA containing the gene encoding either factor VIII or factor IX — with nanoparticles made of chitosan, a tough polymeric carbohydrate found in the exoskeleton of crustaceans. Chitosan protects the therapeutic gene product and chaperones it through the gut. "The oral route has significant appeal," says Gonzalo Hortelano, a gene-therapy researcher at McMaster University in Hamilton, Canada. "The key is to achieve a system of delivery that's persistent, effective and completely safe."

Independent studies by Hortelano's group and other research teams in Germany and the United States have shown that this oral gene therapy does not activate the immune system. Indeed, exposure of the protein produced by the nanoparticle-based gene therapy to the gut mucosa prevents inhibitor development and restores clotting-factor activity in mouse models of both haemophilia A[7,8] and B[9]. "This approach really could hold big benefit for patients," says Jörg Schüttrumpf, a transfusion-medicine specialist who led one of the studies performed at the German Red Cross Blood Donor Service in Frankfurt.

Kam Leong, a biomedical engineer at Columbia University in New York City whose team was the first to demonstrate success with this approach in mice[7], has even tried feeding the chitosan–DNA nanoparticles to dogs with haemophilia A. Leong found some evidence of gene transfer and a reduction in inhibitors in the animals. But bleeding times were not reduced, which would be expected if sufficient levels of factor VIII were being produced. "It is still a very inefficient process," Leong says, "so it requires continued optimization."

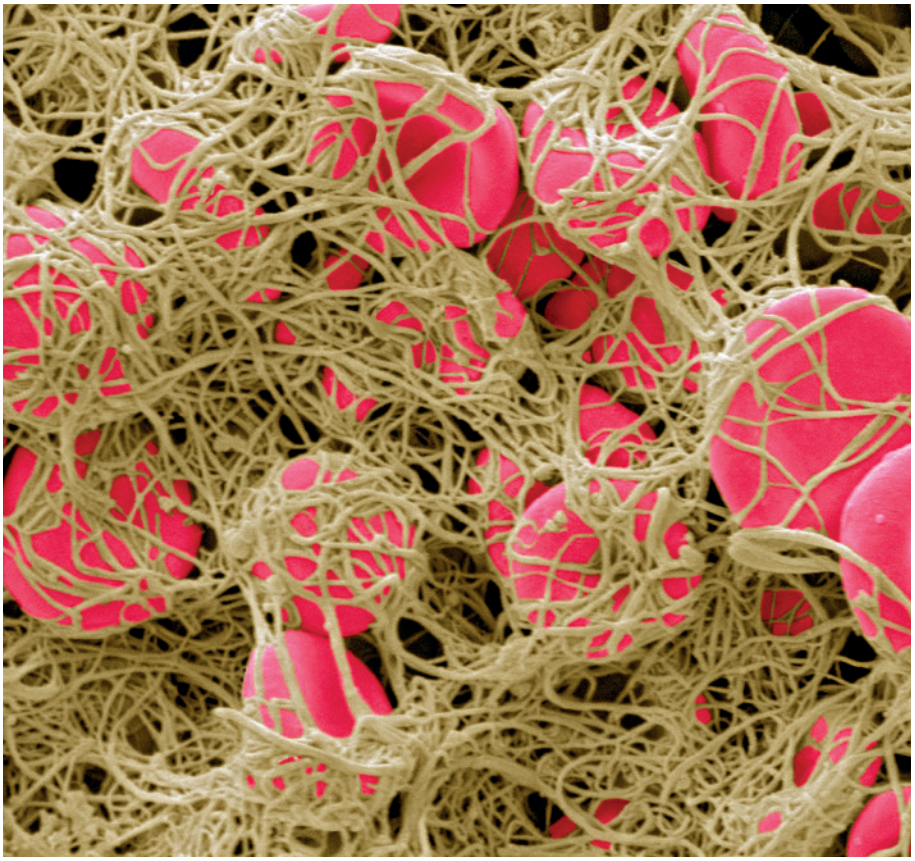
Although the ideal remains a gene therapy that both corrects the disease and offers immune tolerance, some scientists have focused on treating inhibitor formation, without worrying about fixing the disease. Under this strategy, people would still need to take factor-replacement therapies, but they could do so without fear of inhibitor development.

With this in mind, independent teams led by Scott and Herzog took the conventional viral-vector approach to inducing tolerance through gene therapy. But rather than delivering the entire gene for the clotting-factor proteins to cells, as most gene therapies do, the researchers used the viruses to engineer immune-regulating B cells to express a fragment of the clotting factor fused to an immune molecule called an immunoglobulin. This led to long-lived tolerance in mouse models of haemophilia A[10] and B[11].

Pursuing such gene-therapy approaches offers a degree of bet hedging, says Herzog. "Each strategy has potential advantages and disadvantages," he points out, "and we do not really know yet what will work or may work best in people." With so many therapeutic tactics moving through the preclinical pipeline, scientists and clinicians remain hopeful that at least one will ultimately succeed, eliminating the problem of inhibitor formation for people with haemophilia altogether. ■

Elie Dolgin *is a science writer in Somerville, Massachusetts.*

1. Verma, D. *et al. Proc. Natl Acad. Sci. USA* **107,** 7101–7106 (2010).
2. Sherman, A. *et al. Blood* **124,** 1659–1668 (2014).
3. Rawle, F. E. *et al. J. Thromb. Haemost.* **4,** 2172–2179 (2006).
4. Ruhlman, T., Ahangari, R., Devine, A., Samsam, M. & Daniell, H. *Plant Biotechnol. J.* **5,** 495–510 (2007).
5. Zhang, A.-H. A. *et al. Blood* **122,** 2337 (2013).
6. High, K. A. *Blood* **120,** 4482–4487 (2012).
7. Bowman, K., Sarkar, R., Raut, S. & Leong, K. W. *J. Control. Release* **132,** 252–259 (2008).
8. Dhadwar, S. S., Kiernan, J., Wen, J. & Hortelano, G. *J. Thromb. Haemost.* **8,** 2743–2750 (2010).
9. Quade-Lyssy, P. *et al. J. Thromb. Haemost.* **12,** 932–942 (2014).
10. Lei, T. C. & Scott, D. W. *Blood* **105,** 4865–4870 (2005).
11. Wang, X. *et al. Mol. Ther.* **22,** 1139–1150 (2014).



STEVE GSCHMEISSNER/SPL

The control of blood clotting treads a fine line between promotion and inhibition.

THROMBOSIS

Balancing act

A promising therapy curtails clotting inhibitors rather than replacing proteins that promote blood clotting.

BY CASSANDRA WILLYARD

Kanjaksha Ghosh has seen more than a thousand people with haemophilia since he became a physician. But he has always wondered why some patients bleed spontaneously and develop crippling joint damage whereas others barely seem to be affected.

Ghosh, who heads the National Institute of Immunohaematology in Mumbai, India, remembers a soldier who had been fighting insurgents in the northeast of the country. The man's brother was almost bedridden by haemophilia, but the soldier's symptoms were so mild that he did not even realize that he had the disease until he was shot on the battlefield.

In the 1990s, Ghosh began trying to work out why such discrepancies existed by studying families like the soldier's. When he delved into the genomes of those with a milder disease, he often saw not just a mutation in the affected clotting-factor gene, but also a mutation in another gene — the first causing haemophilia, the tendency to bleed, and the second causing thrombophilia, the tendency to clot. Ghosh's research leads to the conclusion that a patient with haemophilia who co-inherits a thrombophilic gene bleeds less than one without that mutation.

Blood coagulation is regulated by one set of proteins that causes clotting and another set that prevents it (see 'Perfect balance'). Too little clotting ability leads to bleeding disorders. Too much leads to vessel-blocking clots that can cause strokes and heart attacks. Existing haemophilia treatments tip the balance towards clotting by adding what the body lacks — the clotting factor that is missing or defective. But natural human experiments such as Ghosh's soldier suggest an alternative strategy to treat the disease. Rather than boosting the factors that promote clotting, researchers might instead disable the anticoagulation machinery that prevents clotting.

In the past few years, three drug companies have moved compounds aimed at inhibiting anticoagulation into clinical trials. The hope is that these therapies will be as effective as existing treatments and much more convenient. Rather than receiving multiple infusions of protein replacement each week, patients might be able to control their bleeding with long-lasting injections.

TARGET PRACTICE

The complex cascade that results in the formation of a clot begins when a blood vessel is injured. Several proteins hold the process in check to prevent clots from forming where they are not needed. One such protein, tissue factor pathway inhibitor (TFPI), impedes the initiation of coagulation. Studies published over the past two decades suggest that blocking this protein can promote clotting, which could curb bleeding in people with haemophilia.

The Danish pharmaceutical company Novo Nordisk in Bagsvaerd began working on an antibody designed to inhibit TFPI in the 1990s. Its researchers showed that this antibody could speed up clot formation in blood plasma from people with haemophilia[1]. They also found that it could shorten bleeding time and hasten clotting in rabbits with induced haemophilia. These results seemed promising, but Novo Nordisk began pursuing other strategies to treat haemophilia, and research to develop an anti-TFPI antibody was halted.

In 2006, Novo Nordisk decided to look for therapies that could be injected under the skin and revived the programme. By 2010, the company had launched a clinical trial in Europe and Asia to test the safety of an anti-TFPI monoclonal antibody called concizumab. The researchers administered the antibody either intravenously or subcutaneously to 28 healthy volunteers and 24 people with haemophilia. Preliminary results presented in 2013 at the International Society on Thrombosis and Haemostasis meeting in Amsterdam suggest that concizumab is safe, and that it can improve coagulation. Participants did not report any severe adverse events, although one of the healthy volunteers in the group receiving the highest dose of concizumab developed a small blood clot that disappeared on its own.

The company hopes to launch a second study in mid-2015 to determine the appropriate dose before moving on to test the efficacy of the treatment. "We have liked TFPI as a target for a long time," says Ida Hilden, scientific director of Novo Nordisk's concizumab project.

Drug company Baxter International, based in Deerfield, Illinois, sells recombinant clotting factors for treating haemophilia and also has its sights on TFPI. In the same year that Novo Nordisk launched its concizumab trial, Baxter struck a deal to purchase a suite of haemophilia-related assets from the former therapeutics company Archemix. Those assets included a therapy designed to inhibit TFPI that had already entered a safety study in the United Kingdom. This therapy was an aptamer, a small strand of nucleotides

designed to inhibit TFPI's activity by binding to it, much like an antibody.

The compound, known as BAX 499, performed well in animal studies but failed to deliver in humans[2]. In 2012, Baxter halted the trial due to an increased number of bleeding events. The failure came as a shock. "We did extensive safety studies in monkeys," says Fritz Scheiflinger, vice-president of research and innovation at Baxter BioScience in Vienna. "We gave huge amounts of aptamer over six months", yet there were no signs that the compound was unsafe, he says.

Scheiflinger and his colleagues think that they now have an explanation for this strange effect. TFPI lasts no more than a couple of hours in the bloodstream, but BAX 499 has a longer half-life. When BAX 499 binds to TFPI, it allows the protein to persist for longer and, over time, to accumulate. And although the drug binds to TFPI, it does not completely deactivate it. So, as partially active TFPI piles up, the balance eventually tips from a pro-clotting effect to an anti-clotting effect. The problem seems to be confined to this particular compound, but nonetheless, the company has shifted its focus away from aptamers.

Baxter is now concentrating on peptides — short strings of amino acids that can be tailored to block part of the TFPI protein — a strategy that Scheiflinger and his colleagues first considered in 2005. The company has identified several promising candidates, but has not yet decided whether it will move them into clinical trials.

TFPI is not the only target for companies hoping to hamper the anticoagulant system. Alnylam Pharmaceuticals in Cambridge, Massachusetts, has set its sights on antithrombin — a protein produced by the liver that hinders clotting. "Antithrombin is probably one of the most potent natural anticoagulants we have in the body," says Benny Sorensen, medical director of clinical research and development at Alnylam. But rather than inhibiting antithrombin's activity, the company plans to block its expression by using short strands of RNA to silence the messenger RNA that carries the code for antithrombin — an approach called RNA interference.

The company is testing its therapy, called ALN-AT3, in a safety study, and the initial results were presented at the World Federation of Haemophilia annual meeting in Melbourne, Australia, in May. After giving healthy volunteers a single low dose of the drug, expression of antithrombin was reduced by 28–32% — an outcome that Sorensen says left the researchers "very surprised". They had thought that it would take higher doses to achieve such a result.

But Sorensen believes that they can do even better. In that first phase, the researchers were not allowed to exceed a 40% reduction in antithrombin because of the safety risks to healthy volunteers. The next phase of the study will include people with haemophilia, and there will not be the same limitation. So the researchers plan to administer multiple doses of the drug. Sorensen thinks that if they can achieve a 50–80% reduction in antithrombin, ALN-AT3 may be able to control bleeding in people with haemophilia without infusions of clotting factor.

PERFECT BALANCE

The body must maintain a delicate equilibrium to ensure that blood flows freely most of the time but clots when necessary. Haemophilia tips the scale towards bleeding, but researchers are looking for new ways to restore the equilibrium.

HAEMOPHILIA

People with haemophilia do not produce enough factor VIII or factor IX, proteins that play a crucial part in clotting.

FACTOR REPLACEMENT TREATMENT

To prevent and staunch bleeding, physicians typically give patients with haemophilia infusions of the factors they lack. Adding these extra factors restores the balance between bleeding and clotting.

ANTICOAGULANT INHIBITION TREATMENT

An approach under development restores balance instead by inhibiting the proteins that prevent clotting – natural anticoagulants such as tissue factor pathway inhibitor (TFPI) and antithrombin.

CAUTIOUS OPTIMISM

All of these therapies have one major advantage over protein replacement: antibodies, peptides and RNA can be effective even when injected under the skin, in part because they are so much smaller than the proteins used for factor-replacement therapy. Novo Nordisk envisages putting its antibody into a 'pen' like the one that people with diabetes use to administer insulin. This would be much more convenient than the intravenous infusions required for existing therapies. "Haemophilia patients are pestered from when they are one or two years old for the rest of their lives with intravenous injections," Sorensen says. "If we can achieve a correction of this haemostatic imbalance that would prevent spontaneous bleeds, then we've really offered an unbelievable change in the lives of these haemophilia patients."

If compounds such as concizumab and ALN-AT3 prove effective, they will undoubtedly be a boon for at least one group of people with haemophilia: those who develop inhibitory antibodies against the blood-clotting factors VIII and IX, and who can no longer receive this standard therapy. Roughly 5% of those with haemophilia B fall into this category, and 30% of those with haemophilia A (see page S166). Baxter, Novo Nordisk and Alnylam think that their products will appeal to other people with haemophilia. But whether these therapies will be safe and effective enough to replace infusions of clotting factor "is the million-dollar question", Scheiflinger says. Sorensen is the most optimistic. He speculates that a once-a-month dose of ALN-AT3 might control bleeding without the need for prophylactic infusions of clotting factor. Even if patients cannot completely forgo factor replacement, he adds, ALN-AT3 might allow them to use less, which could reduce the risk of developing inhibitors.

But many of the physicians who treat patients with haemophilia are not convinced. "The common thinking among haemophilia treaters is that these new strategies can never replace treatment with factor VIII and IX in non-inhibitor patients," says Erik Berntorp, a haematologist at Lund University in Malmö, Sweden. David Ginsburg, a geneticist at the University of Michigan, Ann Arbor, is equally cautious. "In the case of a genetic deficiency, it's pretty hard to improve on replacing the missing factor," he says.

Kenneth Mann, a biochemist at the University of Vermont in Burlington, does not doubt that blocking these anticoagulant pathways will increase the production of thrombin, a key protein in clotting, but he does not think that these therapies will necessarily work for everyone. People with haemophilia "are more heterogeneous than we'd like to admit," he says. And companies will have to work out how to stratify patients on the basis of their real bleeding risk to determine who will benefit from these new approaches. "I don't mean to throw a wet blanket on this," he says, "but caution is required."

One risk is that these therapies will work too well, tipping the balance towards clotting. In a person without haemophilia, Ginsburg says, a total lack of antithrombin "seems to be disastrous". Mice that lack either antithrombin or TFPI die *in utero*. Although the antithrombin-based therapies for haemophilia are not designed to completely block their targets, "knocking them down is not without risk", he says. And as the failure of BAX 499 shows, the risks posed by any new medication can be hard to predict.

Jakob Back, vice-president of the concizumab project at Novo Nordisk, understands the scepticism. Protein replacement has been the go-to therapy for haemophilia for decades. Concizumab and similar therapies represent "a completely different way of approaching haemophilia compared to anything we've been doing for the last 50 years", he says. "We are moving into unknown territory." ■

Cassandra Willyard *is a freelance writer in Madison, Wisconsin.*

1. Nordfang, O., Valentin, S., Beck, T. C. & Hedner, U. *Thromb. Haemost.* **66,** 464–467 (1991).
2. https://ash.confex.com/ash/2012/webprogram/Paper51012.html

JESSICA FORTNER

ORTHOPAEDICS

Joint effort

The hunt is on for ways to diagnose and treat the joint problems that are now the main chronic problem in haemophilia.

BY KATHARINE GAMMON

As a physician who cares for adults with haemophilia, Annette von Drygalski sees patient after patient with bulging, painful knees and elbows caused by bleeding into the joint. The rise in cases of this crippling condition, which can lead to arthritis and disability, drives the work of von Drygalski and her team at the University of California's San Diego Medical Center — part of a growing body of researchers studying haemophilic joint disease and the pain that it causes.

Before clotting factor became widely available as a treatment (see page S162), people with haemophilia rarely reached adulthood, so haemophilic joint disease was not on the radar of most research programmes. But now that people with the disease have a life expectancy similar to that of the general population, arthritis caused by the disorder has emerged as a serious medical problem.

A bleed inside a joint leads quickly to stiffness and pain. The residual iron from pooled blood causes inflammation of the joint lining, a condition known as synovitis. Physicians can remove the inflamed tissue surgically (which, for people with haemophilia, comes with a high risk of bleeding) or by injecting radioisotopes into the joint. These emit radioactive particles that destroy the cells in the joint lining and prevent further bleeding. Such surgeries are delicate procedures, says Mauricio Silva, an orthopaedic surgeon at the University of California, Los Angeles, who specializes in haemophilic joint operations. "The deformities are much more severe than someone with arthritis," he says.

The basic remedy for bleeding into the joint has been for patients to self-administer more clotting factor when they believe they are having a bleeding episode. But this is expensive, and does not help everyone. "This field will require lots of new thoughts, beyond administering clotting factor for joint health, over the next decade to improve the life of those with haemophilia," says von Drygalski.

Researchers are tackling the problem from multiple directions: through better imaging, by using novel biomarkers that might be able to reveal even minor joint bleeds, and by applying knowledge from other types of arthritis. It will take research in all of these areas to work out new ways to diagnose and treat haemophilic joint disease and understand its causes.

JOINT INSPECTION

One problem is that there is no definitive way for physicians to distinguish between normal arthritic joint pain and that caused by a bleed. Von Drygalski's research shows that only one-third of painful episodes reported by people with haemophilia are associated with bleeding into the joint[1]. Similarly, physicians find it hard to determine the cause of joint pain: in one small study[1], von Drygalski and her colleagues found that physicians' assessments, based on patient interviews and physical examinations, were incorrect in 18 of 40 instances.

Imaging technologies can help. The highest-quality pictures come from magnetic resonance imaging (MRI), but these systems are slow, bulky and costly to run, and so are not commonly used in haemophilia clinics.

With an eye on those drawbacks, von Drygalski and her colleagues developed a clinical tool that uses ultrasound. The musculoskeletal ultrasound (MSKUS) system — featuring a hockey-stick-shaped ultrasound probe — can distinguish between bleeding and inflammation during painful episodes. As part of a large initiative in Europe sponsored by pharmaceutical giant Pfizer, staff at about 10–15 haemophilia treatment centres are currently being trained to use the technology. The same initiative is in the planning stages in the United States, where training will be given at 10 centres.

MSKUS checks the crevices of joints for inflammation or bleeding, and is less costly than MRI but just as accurate, says von Drygalski. In particular, she says, ultrasound provides greater detail on what is happening in acutely and chronically painful haemophilic joints, where bleeding has caused both synovitis and inflammatory changes to soft tissue.

MIKE DEVLIN/SPL

An X-ray of the knees of a person with haemophilia, both damaged from bleeding inside the joints.

MOLECULAR MARKERS

The molecular basis of how haemophilia results in joint pain is still not clear. One hypothesis is that the blood of patients with the disease is a poor activator of a key protein called thrombin activatable fibrinolysis inhibitor (TAFI), which controls clot stability and reduces inflammation. For example, administering additional TAFI relieves discomfort in non-haemophiliacs with inflammatory arthritis. Because the protein stops blood clots from breaking down, it helps people with haemophilia to form clots and maintain them. Von Drygalski, in collaboration with Laurent Mosnier, an assistant professor of molecular medicine at the Scripps Research Institute in La Jolla, California, is studying how treating patients with extra TAFI might help to relieve haemophilia joint problems.

Mosnier, for his part, is doing basic molecular studies to better understand the contribution of clot breakdown in bleeding, and to investigate whether TAFI can be genetically modified to make it more potent and diminish bleeding complications.

To tease out TAFI's clotting and anti-inflammatory roles — and to find out why TAFI may not be fully functional in people with haemophilia — both researchers are using haemophilic mouse models as well as mice that have been engineered to lack the gene that encodes TAFI. Von Drygalski hopes that this will lead to treatments beyond the standard infusions of clotting factor. If it is established that poor TAFI activation in haemophilia contributes to joint disease and inflammation, researchers could develop engineered versions of TAFI with high potency that persist for longer in the body. The researchers hope that such agents could eventually mitigate or even prevent haemophilic joint disease.

DRUG SEARCH

Ideally, physicians would like to have a test that determines which people with haemophilia have the highest risk of developing joint disease. At Rush University in Chicago, Illinois, molecular biologist Narine Hakobyan has found about half a dozen biomarkers in the blood of haemophilic mice[2] that could signal very minor bleeds before damage occurs in the joint.

She and her colleague Leonard Valentino (who now works at health-care company Baxter International in Deerfield, Illinois) set out to create animal models for haemophilic joint degradation in 2001. They made one mouse model that had joint bleeds after injury and another that bled into the joint even in the absence of trauma. They also created a scoring system to evaluate how well drugs stopped bleeding in the joints, which could be used to rank the effectiveness of new drugs.

Hakobyan's study[2] revealed biomarkers that could be detected after injecting just 25 microlitres of blood into the joints of mice that lack clotting factor — showing that even tiny bleeds have markers that could be used to predict joint deterioration. These could guide scientists' search for new drugs to treat haemophilic joint disease, and could point to the fundamental mechanisms underlying the illness. "It would be helpful to know at which point joint disease is reversible, and where we can act to use drugs as therapeutic agents," says Hakobyan. Other markers are likely to be found for different stages of the disease, Hakobyan says.

BEYOND CLOTTING FACTORS

To better understand the joint and its response to bleeding, researchers are studying changes to the bone around it. This may require creative thinking about mechanisms beyond clotting factors, says Paul Monahan, a haematologist at the University of North Carolina in Chapel Hill, who has studied whether rheumatoid arthritis drugs can improve mobility and reduce inflammation in haemophilic mice.

Monahan thinks that treatment with infusions of clotting factor, known as prophylaxis, is not a good way to treat all patients with haemophilia, especially those who have breakthrough bleeding — bleeds that happen in between their infusions of clotting factor. For instance, previous research[3] has shown that regularly giving extra doses of clotting agent beyond what is needed for primary prophylaxis adequately controls joint bleeding in less than 40% of people with haemophilia.

He likens this approach to giving only one therapy to patients with asthma. "You wouldn't treat an asthmatic with just a bronchodilator — you need to address both the acute spasm and the underlying inflammation," he says. Likewise, patients with haemophilia could potentially be treated with drugs that reduce inflammation as well as being given clotting factor.

Another potential therapy is the use of special radioisotopes to attack the inflamed joint lining. In July, Navidea Biopharmaceuticals of Dublin, Ohio, announced a partnership with the start-up firm Rheumco to develop a tin radioisotope technology that blasts out inflamed joint tissue. The idea is to inject a colloidal suspension of tin-117 particles into the joints of children with haemophilia. This radioisotope was selected because it has a small, focused area of radiative impact, so there is less chance of radiation damaging nearby tissue — an important consideration for children whose bones are still growing.

Navidea and Rheumco are completing animal testing for the tin-isotope project and are optimizing the technology for use in people. Being able to treat children with the method would be a boon because early treatment is key for these disorders, says Mark Pykett, formerly chief executive of Navidea and now chief executive of Agilis Biotherapeutics in New York. Physicians have identified joint microbleeds in patients as young as two years old. "If you can prevent that, 10 or 20 years down the road, they will be better off," he says.

The limited treatment options for haemophilic children and adults with joint pain strongly motivates researchers. Only a few decades ago, patients with haemophilia did not have the chance to grow old; now they are feeling the effects of living for longer with the disease. "Joints are so important," says von Drygalski, "because people are living to 60 or 70 years old — just trying to live normal lives." ■

Katharine Gammon *is a freelance science writer in Santa Monica, California.*

1. Ceponis, A., Wong-Sefidan, I., Glass, C. S. & von Drygalski, A. *Haemophilia* **19,** 790–798 (2013).
2. Hakobyan, N. *et al. J. Thromb. Haemost.* **12** (suppl. 1), 1 (2014).
3. Greene, W. B., McMillan, C. W. & Warren, M. W. *Clin. Orthop. Relat. Res.* **343,** 19–24 (1997).

T. C. NICHOLS, B. J. Q. NICHOLS & S. A. GALLAGHER

Haemophilic dogs at the University of North Carolina's blood-research laboratory are helping researchers to learn about the disease and develop treatments.

ANIMAL MODELS

Dogged pursuit

In the study of haemophilia, man really does have a best friend.

BY EMILY SOHN

Austin, a fluffy white-and-black Old English sheepdog, was still a puppy when his owners called the University of North Carolina's Francis Owen Blood Research Laboratory in Chapel Hill four years ago. After deciding that the children were finally old enough to get a dog, the family had quickly bonded with the rambunctious pup. But within six months of bringing Austin home, they had spent US$10,000 on veterinary bills to deal with extreme bleeding from small scrapes. Austin was also suffering from spontaneous bleeding into his joints and uncontrollable nosebleeds caused simply by overexcitement. The family loved him, but could not take care of him.

Timothy Nichols, director of the North Carolina lab, gets enquiries about haemophilic dogs from around the world four or five times a year. Sometimes he offers advice and information. Other times, he goes and gets the dog. After blood tests confirmed that Austin had haemophilia, two of Nichols' lab members flew to the family's home in New Orleans, Louisiana, where they rented a car, packed it with a cool box full of medication and drove Austin back to Chapel Hill. There, the dog joined a colony that for nearly seven decades has been quietly transforming understanding of haemophilia.

Unlike the rats favoured as animal models for many other diseases, dogs develop haemophilia naturally, have enough blood to contribute to research studies and live long enough to reveal long-term outcomes of treatments. "We have a 60-year track record now showing that if it works well in dogs, it's likely going to work well in humans," says Nichols.

LIKE HUMAN, LIKE DOG

The earliest recognized cases of haemophilia in dogs were documented in 1935 in three related Scottish terriers. About a decade later, a lawyer in New York contacted the North Carolina blood-research lab to discuss two Irish setters that were bleeding frequently, both inside and out. Already eager to acquire an animal model of haemophilia, the lab's then-director, Kenneth Brinkhous, adopted the aristocratic, long-haired dogs and began searching for breeding partners for them. Since then, colonies of haemophilic dogs have sprung up at Queen's University in Kingston, Canada; the University of Alabama at Birmingham; and Nara University in Japan. There are also a few dogs at Cornell University in Ithaca, New York. Today, these colonies breed both haemophilic and healthy dogs to maintain populations with specific variants of the disease.

It did not take long for dogs to become pivotal to scientists' understanding of the disorder in humans: the disease works in the same way in both species. Early breeding efforts in the 1940s, for example, made it clear that in dogs, the genes responsible for haemophilia lie on the X chromosome — which later proved to be true for people, too. Except in rare cases, only males get the disease; females are carriers. "The genetic and laboratory studies from breeding these dogs and testing their blood helped establish the classic parallel example of humans

and animals having the same genetic defects," says Jean Dodds, a veterinary surgeon in Santa Monica, California, who has been working with haemophilic dogs since 1959.

More recently, gene-sequencing studies have revealed that identical genes with parallel mutations account for many cases of the disease in both dogs and humans. Both species can have either haemophilia A or haemophilia B, versions of the condition caused by defects in the genes that produce the clotting proteins factor VIII and factor IX, respectively. Symptoms are remarkably similar across species: both people and dogs with the disease are unable to form clots, so cuts can bleed uncontrollably. Bleeding in the bowel can lead to diarrhoea. And lumps of blood can form in joints and muscles.

Dogs are also good models for practical reasons. Most of them are bigger than small children. They react to medicines much like humans do, allowing researchers to look to dogs first as they calculate doses. And the animals cooperate well. "The dogs here are around people all the time," says Nichols. "If you need to draw blood, they put their paws out."

DOGS FIRST

Dogs in haemophilia colonies often win researchers' hearts. Veterinary surgeon Clint Lothrop of the University of Alabama at Birmingham has adopted several from his colony, and he treats them at home when they bleed. The Queen's University dogs run, climb and play with balls and other toys every afternoon, says Queen's haematologist David Lillicrap. The North Carolina dogs have access to an outdoor play area. With severe haemophilia, animals can bleed simply from wearing collars, so handlers are careful to prevent fights or rough play.

Between play sessions, dogs give blood for research. Those donations have allowed scientists to make key discoveries about why the disease develops.

By the 1950s, researchers knew that normal blood could correct clotting defects, but they were not sure which components of blood mattered most. With the help of dog blood, Brinkhous and others deduced[1] that clotting factors were in the plasma rather than mixed in with platelets or blood cells. Giving healthy plasma to haemophilic dogs made them better. Once scientists had identified factors VIII and IX, and could distinguish between healthy and haemophilic dogs, Brinkhous and his colleagues were able to develop a test for measuring levels of the factors in plasma on the basis of how long it took for clots to form in test tubes.

In the 1940s, life expectancy for humans with haemophilia had been about 20 years, often plagued by painful bleeding into muscles and joints, says Nichols. Plasma-replacement therapy transformed the quality — and duration — of life, as did the ability to concentrate the factor in plasma, developed by the mid-1960s.

In the 1970s and early 1980s haemophilia treatment went through a dark period: contaminated plasma infected many recipients with hepatitis or HIV. Dogs helped people out of this tragic stage.

Scientists thought that they had found the light at the end of the tunnel in 1984, when the cloning of the gene for factor VIII allowed them to make artificial factor in the lab[2]. But after years of dealing with blood-borne infections and a cultural fear of such genetically modified products, it was hard to get people to try the synthetic factor. Then studies[3] in dogs showed that the treatment worked without complications, and a 43-year-old North Carolina state legislator agreed to be the first person to sign up. "He knew of Brinkhous's work and he knew of the dogs at Chapel Hill and it helped him to know that it had really helped the dogs and was safe," says Nichols.

To everyone's relief, the treatment worked. In fact, infusion of the factor was so uneventful that the recipient, known as GM, pretended to be a hamster during the procedure (the product had been produced in hamster cells). After the treatment was licensed, GM spoke at a celebration at the Genetics Institute in Cambridge, Massachusetts. "After slowly and painfully climbing to a balcony half way up the stairs, he delivered a powerful story about what it was like to grow up with hemophilia without adequate treatment, how as a child he had lost a beloved older brother from a bleed, and how important the development of safe recombinant factors was to him and all people with hemophilia," wrote Gilbert White, director of the Blood Research Institute at the BloodCenter of Wisconsin in Milwaukee, in a paper[4] describing 35 years of advances in haemophilia research. "His comments had the entire company in tears."

> ***"We have a 60-year track record showing that if it works well in dogs, it's likely going to work well in humans."***

POINTING THE WAY

Research in canines often foreshadows what is coming for humans. Over the years, more than 25 products that had been tested in dogs have been licensed for clinical use in people. One of the first studies to show the feasibility of gene therapy[5], published in 1993, involved three factor-IX-deficient dogs and an extremely invasive procedure, in which researchers removed two-thirds of each dog's liver. Over the course of three days, they injected the regenerating organs with a potentially dangerous HIV-like virus loaded with the healthy gene. The procedure boosted levels of factor IX from zero to 1% of normal — enough to fuel optimism that a more efficient procedure might one day be possible.

By 1999, dog studies[6] began to show that one injection with a much safer vector called an adeno-associated virus could deliver a healthy factor IX gene, boosting levels of the clotting factor to 2% — enough to reduce spontaneous bleeds. "We were able to move past that rapidly and have had levels of 10% for a long time," says Katherine High, a haematologist at the Perelman School of Medicine at the University of Pennsylvania in Philadelphia. Dogs can now get simple, 10- to 15-minute infusions of factor-bearing viral vectors. Similar work with factor VIII is close behind, says Nichols.

Some of the first dogs to receive factor IX gene therapy with just a single injection have lived full and happy lives. Brad and Semi were two basenjis — African hunting dogs — who lived in the Alabama colony. After receiving the treatment, one died at 13, the other at 14, neither from haemophilia-related causes. Several clinical trials are now assessing gene therapy with factor IX in humans (see page S160).

Other studies are testing the possibility of administering factors VIII and IX orally instead of with an injection — a technique that has been shown to work in mice and is now being tested in dogs. And ongoing work by Lothrop and his colleagues suggests that replacement factors might become available as longer-lasting, less-invasive subcutaneous shots instead of intravenous injections.

Dogs are also helping scientists to develop strategies for combating the inhibitor antibodies that many patients develop in response to factor-replacement therapy. One approach[7] gives dogs a gene to express another clotting factor, factor VIIa, completely bypassing the need for factors VIII and IX. The technique can reduce the number of bad bleeds each year from five or ten to one or even none.

In other lines of work, dogs have undergone bone-marrow transplants to express factor VIII in their platelets, shielding them from inhibitors. And Nichols' team has acquired a strain of dogs deficient in clotting factor VII, allowing it to test therapies for rare bleeding disorders that may not occur in enough humans to allow large clinical trials.

It is unlikely that any of these next-generation approaches would have been possible without canine models. "The role of haemophilic dogs in the preclinical development of novel therapies for haemophilia during the past three decades has been enormous," says Lillicrap. The disease once seemed insurmountable, but in the years ahead, he says, dogs will continue to provide insights that will make life better for humans. ■

Emily Sohn *is a freelance journalist in Minneapolis, Minnesota.*

1. Brinkhous, K. M., Penick, G. D., Langdell, R. D., Wagner, R. H. & Graham, J. B. *AMA Arch. Pathol.* **61,** 6–10 (1956).
2. Wood, W. I. *et al. Nature* **312,** 330–337 (1984).
3. Giles, A. R. *et al. Blood* **72,** 335–339 (1988).
4. White, G. C. *Trans. Am. Clin. Climatol. Assoc.* **121,** 61–75 (2010).
5. Kay, M. A. *et al. Science* **262,** 117–119 (1993).
6. Herzog, R. W. *et al. Nature Med.* **5,** 56–63 (1999).
7. Margaritis, P. *et al. Blood* **113,** 3682–3689 (2009).